6-14-2019

Estimation for Zero-Inflated Beta-Binomial Regression Model with Missing Response and Covariate Measurement Error

Rong Luo
University of Windsor

Estimation for Zero-Inflated Beta-Binomial Regression Model with Missing Response and Covariate Measurement Error

by

Rong Luo

A Dissertation
Submitted to the Faculty of Graduate Studies
through the Department of Mathematics and Statistics
in Partial Fulfilment of the Requirements for
the Degree of Doctor of Philosophy
at the University of Windsor

Windsor, Ontario, Canada

2019

Estimation for Zero-Inflated Beta-Binomial Regression Model

with Missing Response and Covariate Measurement Error

by

Rong Luo

APPROVED BY:

N. Klar, External Examiner
Western University

Y. Aneja
Odette School of Business

M. Hlynka
Department of Mathematics and Statistics

M. Belalia
Department of Mathematics and Statistics

S. R. Paul, Advisor
Department of Mathematics and Statistics

June 14, 2019

Declaration of Co-Authorship / Previous Publication

I. Co-Authorship

I hereby declare that this dissertation incorporates material that is result of joint research, as follows: Chapter 3 of the dissertation was co-authored with Dr. Sudhir Paul. In all cases, the key ideas, data analysis, interpretation, and writing were performed by the author. Dr. Sudhir Paul provided feedback on refinement of ideas and editing of the manuscript.

I am aware of the University of Windsor Senate Policy on Authorship and I certified that I have properly acknowledged the contribution of the other researchers to my thesis, and have obtained written permission from each of the co-author(s) to include the above material(s) in my thesis.

I certify that, with the above qualification, this dissertation, and the research to which it refers, is the product of my own work.

II. Previous Publication

This dissertation includes one original paper that has been previously published, as follows:

Dissertation Chapter	Publication title/full citation	Publication status
Chapter 3	Rong Luo and Sudhir Paul, (2018). Estimation for zero-inflated beta-binomial regression model with missing response data. *Statistics in Medicine*	Published

I certify that I have obtained a written permission from the copyright owner(s) to include the above published material(s) in my dissertation. I certify that the above material describes work completed during my registration as a graduate student at the University of Windsor.

III. General

I certify that, to the best of my knowledge, my dissertation does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my dissertation, published or otherwise, are fully acknowledged in accordance with standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained written permission from the copyright owner(s) to include such material(s) in my dissertation and have included copies of such copyright clearances to the appendix.

I declare that this is a true copy of my dissertation, including any final revisions, as approved by my dissertation committee and the Graduate Studies Office, and that this dissertation has not been submitted for a higher degree to any other University or Institution.

Abstract

Discrete, binary data with over-dispersion and zero-inflation can arise in toxicology and other similar fields. In studies where the litter is an experimental unit, there is a "litter effect" which means that the litter mates respond more alike than animals from other litters. In experimental data, foetuses in the same litter have similar responses to the treatment. The probability of "success" may not remain constant throughout the litters. In regression analysis of such data another problem that may arise in practice is that some responses may be missing or/and some covariates may have measurement error. In this dissertation we develop an estimation procedure for the parameters of a zero-inflated over-dispersed binomial model in the presence of missing responses without/with considering covariate measurement errors. A weighted expectation maximization algorithm is used for the maximum likelihood (ML) estimation of the parameters involved. Extensive simulations are conducted to study the properties of the estimates in terms of average estimates (AE), relative bias (RB), variance (VAR), mean squared error (MSE) and coverage probability (CP) of estimates. Simulations show much superior properties of the estimates obtained using the weighted expectation maximization algorithm. Some illustrative examples and a discussion are given.

Dedication

Dedicated to my family

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor Dr. Sudhir R. Paul. His continuous supervision, encouragement, guidance and support make the research presented in this dissertation possible. His rigorous and conscientious attitude, and enthusiasm for research and teaching always inspire me deeply. All the things I learned from Dr. Paul will benefit my future academic and career.

I am very thankful to Dr. Myron Hlynka, Dr. Mohamed Belalia and Dr.Yash Aneja for being part of my Doctoral committee, and more importantly for their critical review and constructive suggestions and comments, which helped me to improve this dissertation. I would like to express my sincere gratitude to the external examiner of my dissertation Dr. Neil Klar, Department of Epidemiology and Biostatistics, Western University, for his critical review, constructive suggestions and comments.

I am very thankful to the Department of Mathematics and Statistics, University of Windsor for providing me with the financial support such as Graduate Assistantships, Ontario Graduate Scholarship (OGS).

Finally, I am forever grateful to my family and friends for their care and help.

Contents

List of Tables

List of Abbreviations

AE	Average estimates
CC	Complete case analysis
CP	Coverage probability
EM	Expectation maximization algorithm
MAR	Missing at random
MC	Monte Carlo method
MCAR	Missing completely at random
MI	Multiple imputation
ML	Maximum likelihood
MNAR	Missing not at random
MSE	Mean squared error
RB	Relative bias
VAR	Variance

Chapter 1

Introduction

Discrete data in the form of proportions can arise in toxicology (Kleinman (1973); Weil (1970)) and other similar fields (Crowder (1978); Donovan et al. (1994); Gibson and Austin (1996); Otake and Prentice (1984)). In studies where the litter is an experimental unit, there is a "litter effect" which means that the littermates respond more alike than animals from other litters. In experimental data, foetuses in the same litter have similar responses to the treatment. The probability of "success" may not remain constant throughout the litters. This effect of litter is known as "heritability of a dichotomous trait" (Elston (1977)) or intra-litter or intra-class correlation. A number of parametric (Skellam (1948); Haseman and Kupper (1979); Altham (1978)) and semi-parametric models (McCullagh (1983); Nelder and Pregibon (1987); Godambe and Thompson (1989)) have been used to analyse this kind of data in the form of proportions. A popular parametric model is the two parameter beta-binomial model, proposed originally by Williams (1975) and later applied by Paul (1982) assuming that the binomial parameter varies between litters. A second problem for discrete data in the form of proportions is that the zero count occurs more often than can be accom-

modated by a binomial model or a beta-binomial model (Johnson et al. (2005)), So, a zero-inflated beta-binomial model might be more appropriate for these data (Deng and Paul (2005)).

A further complication that can arise in practical experimental data analytic situations is that some of the binomial or beta-binomial responses might be missing. A lot of work has been done for the estimation of the parameters for normally distributed data (Little and Rubin (2014); Rubin (1977)) and data that follow generalized linear models with missing data (Ibrahim et al. (2005)). Ibrahim (1990) proposes the method of weights for parameter estimation in incomplete data in a generalized linear model where the missing data has a range. Ibrahim and Lipsitz (1996) use the same method to estimate regression coefficients in a binomial regression model when the nonresponse is nonignorable. Troxel et al. (1997) consider a weighted estimating equation to analyse data with nonignorable missing response. Wang (1999) suggests modified estimating functions to analyse the binary outcome potentially observed at successive time points. Ibrahim et al. (2001) discuss the maximum likelihood estimation method in a generalized linear mixed model when the nonresponse is nonignorable. Stubbendick and Ibrahim (2003) use the maximum likelihood method for nonignorable missing response and covariates in a random effects model. More recently, Mian and Paul (2016) develop procedures for the estimation of the parameters of a zero-inflated negative binomial model with missing values.

Additional complications can arise in practical experimental data analytic situations when one or more of the covariates is measured with error. Measurement error can happen when there is a difference between a measured value of a quantity and its true value. When covariates are measured with errors, the usual regression estimates by using the observed value of covariates, are biased (Stefanski and Carroll (1985)).

Covariate measurement error has been considered to be an important subject in many application areas. For example, in the field of medicine and epidemiology, individual exposure to certain radiation or blood pressure of participants are recorded and the influence on disease is investigated. In the mutagenic study (Lüning et al. (1966)), all individuals in a small group are given the same dose. However, because of the size of the animals the actual dose will vary from animal to animal. In the Framingham study (Kannel et al. (1986)), it is impossible to measure long-term systolic blood pressure. As a substitute, the blood pressure observed during a clinic visit is available. The reason that the long-term blood pressure and single-visit blood pressure differ is that blood pressure has major daily, as well as seasonal, variation (Carroll et al. (2006)).

Many studies about measurement error models have focused primarily on linear models. Adcock (1878) deals with estimation in models of univariate regression including measurement errors in variables. Gleser (1981) considers a multivariate regression model with measurement error in variables. Interest in generalized nonlinear models is also popular. Prentice (1982) proposes an estimation method in Cox's failure time regression model when the regression vector is subject to measurement error. Wolter and Fuller (1982) present an estimation procedure for the coefficients of a nonlinear functional relation, where observations are subject to measurement error. Carroll et al. (1984) consider binary regression models when some of the predictors are measured with error. Stefanski and Carroll (1985) introduce a bias-adjusted estimator and two estimators appropriate for normally distributed measurement errors for a logistic regression model when covariates are subject to measurement error. Schafer (1987) develops the EM algorithm to obtain estimators of regression coefficients for generalized linear models with canonical link when nor-

mally distributed covariates are masked by normally distributed measurement errors. Burr (1988) considers the Berkson case of the errors in variables in a binary regression model. Generalized linear models with covariate measurement error can be estimated by maximum likelihood using gllamm, a program that fits a large class of multilevel latent variable models (Rabe-Hesketh et al. (2004)).

The purpose of this dissertation is to develop inference procedures for the parameters of a zero-inflated beta-binomial regression model where information on some of the covariates are recorded with errors and/or some observations of the binomial responses may be missing. A weighted expectation maximization algorithm (Dempster et al. (1977)) is developed for the maximum likelihood (ML) estimation of the parameters involved.

1.1 Organization of the dissertation

In Chapter 2, we review some literature related to zero inflated over dispersed binary data, missing values issues and the measurement error process. In Chapter 3, we develop an estimation procedure for the parameters of a zero-inflated beta-binomial regression model in presence of missing values in the response variable. Results of a simulation study with an illustrative example and discussion leading to some conclusions is given. Chapter 4 shows the estimation procedure for the parameters of a zero-inflated beta-binomial regression model in presence of measurement error in covariates without/with missing responses. Results of a simulation study with an illustrative example and a discussion leading to some conclusions is given.

A plan for future study is given in Chapter 5. There is repetition in the chapters because the chapters are intended for publication.

Chapter 2

Preliminaries and Literature Review

2.1 Zero-inflated beta-binomial distribution

2.1.1 Binomial data model and beta-binomial Distribution

Suppose Z be a m-dimensional vector of Bernoulli-distributed outcomes, with success probability p. Assuming the elements in Z to be independent given p, then $Y = \sum_{j=1}^{m} Z_j$, conditionally on p has a binomial distribution with parameters n and success probability p. We have

$$P(Y = y|p) = \binom{m}{y} p^y (1-p)^{m-y} \qquad y = 0, \cdots, m.$$

The beta-binomial model (Skellam (1948); Kleinman (1973)) assumes the parameter $p(0 < p < 1)$ to be sampled from a beta distribution with parameters α and β, i.e., the density of p is

$$f(p|\alpha, \beta) = \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)},$$

where $B(\alpha, \beta)$ denotes the beta function. The marginal density of Y is then given by

$$\begin{aligned} f(y|\alpha,\beta) &= \int \binom{m}{y} p^y (1-p)^{m-y} f(p|\alpha,\beta) dp \\ &= \binom{m}{y} \frac{B(y+\alpha, m-y+\beta)}{B(\alpha,\beta)} \\ &= \binom{m}{y} \frac{\Gamma(y+\alpha)\Gamma(m+\beta-y)\Gamma(\alpha+\beta)}{\Gamma(m+\alpha+\beta)\Gamma(\alpha)\Gamma(\beta)}. \end{aligned} \tag{2.1}$$

This is called the beta-binomial distribution with parameters α and β. It can be easily shown that

$$E(Y) = m\left(\frac{\alpha}{\alpha+\beta}\right) \qquad Var(Y) = m\left(\frac{\alpha}{\alpha+\beta}\right)\left(\frac{m+\alpha+\beta}{1+\alpha+\beta}\right).$$

If $\pi = \frac{\alpha}{\alpha+\beta}$ and $\phi = \frac{1}{\alpha+\beta}$, then

$$E(Y) = m\pi \qquad Var(Y) = m\pi(1-\pi)\left[1 + \frac{(m-1)\phi}{1+\phi}\right] = m\pi(1-\pi)\sigma^2,$$

where $\sigma^2 = 1 + \frac{(m-1)\phi}{1+\phi}$. Because $\alpha > 0, \beta > 0$ and $\phi \geq 0$, we have $\sigma^2 \geq 1$ and $Var(Y) \geq m\pi(1-\pi)$. When $\phi \to 0$, the beta-binomial distribution $BB(\pi, \phi)$ tends to the binomial(π) distribution.

2.1.2 Zero-inflated beta-binomial model

When we use the beta-binomial model to analyze over dispersion discrete data, sometimes more zeros are observed than expected. These data can be analyzed as a zero inflated beta-binomial model with probability density function given by

$$f(y|x_i;\alpha,\beta,\omega) = \begin{cases} \omega + (1-\omega)f(0|\alpha,\beta) & \text{if } y = 0, \\ (1-\omega)f(y|\alpha,\beta) & \text{if } y > 0, \end{cases}$$

where ω is zero-inflated parameter and $f(y|\alpha,\beta)$ is defined by (2.1).

if $\pi = \frac{\alpha}{\alpha+\beta}$ and $\phi = \frac{1}{\alpha+\beta}$, then

$$f(y|x;\pi,\phi,\omega) = \begin{cases} \omega + (1-\omega)\dfrac{\prod_{r=0}^{m-1}(1+r\phi-\pi)}{\prod_{r=0}^{m-1}(1+r\phi)} & \text{if } y = 0, \\ (1-\omega)\binom{m}{y}\dfrac{\prod_{r=0}^{y-1}(\pi+r\phi)\prod_{r=0}^{m-y-1}(1-\pi+r\phi)}{\prod_{r=0}^{m-1}(1+r\phi)} & \text{if } y > 0, \end{cases}$$

with $E(Y) = (1-\omega)m\pi$, and $Var(Y) = (1-\omega)m\pi(1-\pi)\frac{1+m\phi}{1+\phi} + (1-\omega)\omega m^2\pi^2$, where ω is the zero-inflation parameter. We denote this distribution by $ZIBB(\pi,\phi,\omega)$ distribution. Inference regarding the parameters of the beta-binomial model and that of the zero-inflated beta-binomial model has been developed earlier (Dean (1992); Deng and Paul (2000)).

2.2 Missing data issue

Missing data make the parameter estimation and inference much more complicated because almost all standard statistical methods are developed based on complete information for all the variables included in the analysis. Absent observations on some variables may make the parameter estimates biased by using the observed information only.

2.2.1 Missing data mechanism

The missing data mechanism is characterized by the relationship between the missingness and the values of the the variables in the data set. Three kinds of missing data mechanism can be identified (Rubin (1977)), which is very useful in practice.

(1). Missing completely at random (MCAR): Suppose variable Y has some missing values. We will say that these values are MCAR if the probability of missing data on Y is unrelated to the value of Y itself or to the values of any other variable in the data set. For example, We want to assess the relationship between the value of people's houses and income . The MCAR assumption would be satisfied if people who did not report their income were unrelated to their house value and income.

(2). Missing at random (MAR)-a weaker assumption than MCAR: The probability of missing data on Y does not depend on the value of Y after controlling for other variables in the analysis (say X). Formally: $P(Ymissing|Y, X) = P(Ymissing|X)$ (Allison (2001)). For example: The MAR assumption would be satisfied if the probability of missing data on income depends on a house value, but is unrelated to income given the house value.

(3). Missing not at random (MNAR): Missing values do depend on the value of unobserved data, perhaps in addition to the observed data, For example: the probability of missing data on income varies according to the house value and income.

The MCAR and MAR mechanisms are ignorable, which means the inference can be done by analyzing the observed data only and without addressing the model of missing data mechanism. In that sense, MNAR is nonignorable. In the nonignorable case, methods that do not model the missing data mechanism are subject to bias. Thus, the missing data mechanism must be modelled to get good estimates of the parameters of interest.

2.2.2 Methods for handling missing data

2.2.2.1 Conventional methods

(1). Complete case analysis: If a case has missing data for one of the variables, then simply delete that case from the analysis. It is usually the default in statistical packages (Briggs et al. (2003)). Advantages: It can be easily used and is most popular with any kind of statistical analysis and no special computational methods are required. Limitations: It can exclude a large fraction of the original sample. It works well when the data are missing completely at random (MCAR), which rarely happens in reality (Nakai and Ke (2011)).

(2). Imputation methods: Substitute each missing value for a reasonable guess, and then carry out the analysis as if there were no missing values. The main imputation techniques are:

(a). Marginal mean imputation: Compute the mean of X using the non-missing values and use it to impute missing values of X.

(b). Conditional mean imputation: Suppose we are estimating a regression model with multiple independent variables. One of them, X, has missing values. We select those cases with complete information and regress X on all the other independent variables. Then, we use the estimated equation to predict X for those cases it is missing.

(c). Hot deck imputation: Replace values from "similar" responding units.

Limitations of imputation techniques in general: They lead to an underestimation of standard errors and, thus, overestimation of test statistics. The main reason is that the imputed values are completely determined by a model applied to the observed data, in other words, they contain no error (Allison (2001)).

2.2.2.2 Advanced methods

(1). Multiple Imputation (MI): The imputed values are draws from a distribution, so they contain some variation. It replaces each missing item with two or more acceptable values, representing a distribution of possibilities (Allison (2001)). The idea of multiple imputation for missing data was first proposed by Rubin (1977).

MI is a simulation-based procedure. Its purpose is not to replace the individual missing values as close as possible to the true ones, but to handle missing data to achieve valid statistical inference (Schafer (1997)).

Limitation of MI method: The condition for the multiple imputation for missing data is that the data should be missing at random (MAR).

(2). The Expectation-maximization (EM) algorithm: It is based on an expectation step and a maximization step, which are repeated several times until the change of estimated parameter reaches a preset threshold. Maximum likelihood estimates are obtained.

The EM algorithm is a general iterative method of maximum likelihood estimation for incomplete data. The essential idea behind the EM algorithm is to calculate the maximum likelihood estimates for the incomplete data problem by using the complete data likelihood instead of the observed likelihood because the observed likelihood might be complicated or numerically infeasible to maximise (Dempster et al. (1977)).

Let y_{obs} be the observed data, y_{mis} be the missing data, R be the missing data indicator, η be the parameters which include the main model parameters and missing data model parameters, and $L(\eta)$ be the complete likelihood of the data. In the E step, at $(t+1)st$ interaction we compute $Q(\eta|\eta^{(t)}) = E(L(\eta|y_{obs}, R; \eta^{(t)})$. In the M step, we obtain $\eta^{(t+1)} = maxQ(\eta|\eta^{(t)})$.

Note that the E step does not always have a closed form. For discrete missing data, we usually apply EM by weighting as following.

$$\begin{aligned} Q(\eta|\eta^{(t)}) = &\quad E(L(\eta|y_{obs}, R; \eta^{(t)}) \\ = &\quad \textstyle\sum_{y\in S_y} lnf(y, R, \eta) \times p(y_{mis} = y|y_{obs}, R; \eta^{(t)}), \end{aligned} \tag{2.2}$$

where S_y is the support of y and $p(y_{mis} = y|y_{obs}, R; \eta^{(t)})$ is called weight. We can denote it as $w^{(t)}$. Then (2.2) can be written as

$$Q(\eta|\eta^{(t)}) = \textstyle\sum_{y\in S_y} lnf(y, R, \eta) \times w^{(t)}.$$

If the missing data come from a continuous variable, we can employ Monte Carlo (MC) method.

2.2.3 Monte Carlo methods

Monte Carlo methods solve the integration problem by sampling and averages. They are a form of stochastic integration used to approximate expectations by invoking the law of large numbers. Suppose we have $x \sim f(x)$ and we want to compute the mean of $g(x)$. We can write

$$\mu = E(g(x)) = \int g(x)f(x)dx.$$

Then the estimate of μ is

$$\hat{\mu}_{mc} = \frac{1}{m}\sum_{i=1}^{m} g(x_i^*),$$

where $x_1^*, \cdots, x_i^*$ are generated from $f(x)$ and $\hat{\mu}_{mc}$ is a Monte Carlo estimate of μ. By the Law of Large Numbers, we have $\hat{\mu}_{mc}$ converges to μ with probability 1 as $m \to \infty$. With this property, if we have an identical and independent sample $x_1^*, \cdots, x_i^*$, we can approximate the expectation of any function with respect to x. Therefore, the integration problem becomes how to get a good sample.

Usually the target distribution $f(x)$ is very complicated and hard to directly sample. We introduce two kinds of common sampling techniques here.

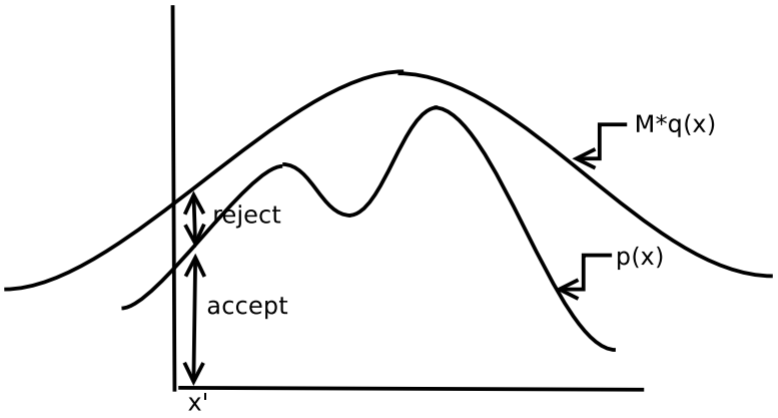
- Importance sampling: We sample from a simpler proposal distribution $h(x)$ instead of $f(x)$ and define the weight function as $w(x) = \frac{f(x)}{h(x)}$. Then we have

$$\mu = E(g(x)) = \int g(x)f(x)dx = \int g(x)w(x)h(x)dx$$

and

$$\hat{\mu}_{mc} = \frac{1}{m}\sum_{i=1}^{m} g(x_i^*)w(x_i^*)$$

- Rejection sampling:



Suppose we want to sample from the density $f(x)$ as shown above. Under most circumstances, it is difficult to sample directly from $f(x)$ if $f(x)$ has a complicated form, for example, multiplication of a few density functions. Rejection sampling is a general method for sampling points independently from a density $f(x)$. In rejection sampling, we can sample in this way:

1. Sample x_i from $h(x)$;
2. Sample u from the uniform distribution $U_{(0,1)}$;

3. Accept the sample x_i if it satisfied that $u < f(x_i)/(Mh(x_i))$; otherwise reject it and return to step 1;

4. Repeat the draws x_i from $h(x)$ until a value is accepted.

M is a constant, finite bound on the likelihood ratio $f(x)/h(x)$, satisfying $1 < M < \infty$. In other words, M must satisfy $f(x) \leqslant Mh(x)$ for all values of x. The main problem with this process is that many samples will get rejected in high-dimensional spaces.

2.2.4 Modelling with missing data

There exist three ways to factor the joint distribution of the complete data and missingness indicators: outcome dependent factorization, pattern-dependent factorization, and parameter-dependent factorization.

For the ith observation, suppose θ is the parameter for main model $f(y_i|x_i)$, while α is the parameter for missingness indicator model $f(r_i|y_i, x_i)$. The three corresponding models available for incomplete data analysis are:

- Selection Model, which factors the joint distribution into a marginal distribution for y_i and a conditional distribution of r_i given y_i, i.e.,

$$f(y_i, r_i \mid x_i, \theta, \alpha) = f(y_i \mid x_i, \theta) f(r_i \mid y_i, x_i, \alpha),$$

  where $f(r_i \mid y_i, x_i, \alpha)$ can be interpreted as self-selection of the ith subject into a specific missingness group.

- Pattern-Mixture Model, which is a pattern-dependent model, and assumes that distribution of repeated measures varies with the missingness patterns and that

the joint distribution is factored as

$$f(y_i, r_i \mid x_i, \theta, \alpha) = f(y_i \mid r_i, x_i, \theta) f(r_i \mid x_i, \alpha).$$

- Shared-Parameter Model. We assume that y_i and r_i are conditionally independent of each other, given a group of parameters ξ_i,

$$f(y_i, r_i \mid x_i, \theta, \alpha) = \int f(y_i \mid \xi_i, x_i, \theta) f(r_i \mid \xi_i, x_i, \alpha) f(\xi_i) d(\xi_i).$$

  Shared parameters ξ_i affect both y_i and r_i, thus can be either observable variables (e.g., gender) or latent variables (e.g., random-effects or latent scores).

2.3 Measurement error process

Measurement error in covariates has three effects (Carroll et al. (2006)):

- It causes bias in parameter estimation for statistical models.
- It leads to a loss of power, sometimes profound, for detecting interesting relationship among variables.
- It masks the features of the data, making graphical model analysis difficult.

In this study we will focus on the first problem. We partition the p dimension vector of covariates x_i for the ith observation as (u_i, z_i), where the vector u_i is observed only indirectly through the measurement w_i and z_i is observed without error. Note that u_i and w_i are q dimensional while z_i is $p - q$ dimensional. The main characteristic of a measurement error issue is that we can observe a variable w_i which is related to u_i and the variable u_i cannot be observed. The parameters in the model relating y_i and (z_i, x_i) cannot be estimated directly by fitting y_i to (z_i, w_i). The goal of parameter

estimation with covariate measurement error is to obtain nearly unbiased estimates of parameters. Substituting w_i for x_i, but making no adjustments in the usual fitting methods for this substitution will lead to estimates that are biased, sometimes very seriously.

2.3.1 Function and structural modelling

(1). Functional modelling: When the unobserved true values are unknown constants (fixed), in which no distribution can be assumed, then the measurement error model is said to be in its functional form.

(2). Structural modelling: When the unobserved true values are identically and independently distributed random variables with mean μ and variance σ^2, the measurement error model is said to be in the structural form.

In this study, we focus on structural measurement error modeling.

2.3.2 Measurement error models

Following Carroll et al. (2006), the measurement error model can be classified into two general types which are used to relate w_i to u_i:

1. Error model, which includes the classical measurement error model.

$$w_i = \tau_0 + \tau_u u_i + \tau_z z_i + e_i.$$

The error term e_i is independent of u_i, z_i and the responses and it is often assumed that e_i has mean zero and it follows a known distribution $f(0, \Sigma)$, where Σ is the covariance matrix of e_i. The intercept τ_0 is a vector, which can be written as $\tau_0 = (\tau_{01}, \cdots \tau_{0q})^T$. The coefficients $\tau_u = (\tau_{u1}, \cdots \tau_{uq})^T$, $\tau_z = (\tau_{z1}, \cdots \tau_{zq})^T$, where $\tau_{uj} (j = 1, \cdots, q)$ and

$\tau_{zk}(k = 1, \cdots, p-q)$ are $q \times 1$ and $(p-q) \times 1$ vectors respectively. If we set $\tau_0 = \mathbf{0}$, $\tau_z = \mathbf{0}$, and $\tau_u = I_q$, where I_q is the $q \times q$ dimensional identity matrix, we have the classical measurement error model.

2. Regression calibration model, which includes the Berkson error model.

$$u_i = \tau_0 + \tau_w w_i + \tau_z z_i + e_i,$$

where, $\tau_w = (\tau_{w1}, \cdots \tau_{wq})^T$. If we set $\tau_0 = \mathbf{0}$, $\tau_z = \mathbf{0}$, and $\tau_w = I_q$, we have the Berkson measurement error model.

2.3.3 Differential and nondifferential Error

Nondifferential measurement error occurs when W has no information about Y given U and Z, which means the measurement error is nondifferential if the distribution of Y given (U, Z, W) depends only on (U, Z), in the other words, Y is conditionally independent of W given the true covariates. For example, we are interested in the long term systolic blood pressure, but we can only measure the blood pressure on a single day. In this situation, a single day's blood pressure value includes no information given by true long term blood pressure. Therefore, that measurement error is nondifferential. Measurement error is differential otherwise. It may happen in case-control studies. For example, in a nutrition study, the outcome is cancer and the true predictor is long term diet before diagnosis, but the reported diet is obtained only after diagnosis. People who develops breast cancer may change their diet, so the reported diet after diagnosis is clearly still related to the cancer outcome.

Under nondifferential measurement error, one can typically estimate parameters in models for responses given the true covariates even though the true covariates are not observed. However, with differential measurement error, one must observe the

true covariate on some observations (Carroll et al. (2006)). In this study, we focus on nondifferential measurement error models.

2.4 Measurement error models and missing data

The typical explanation for the missing data problem (Little and Rubin (2014)) is that values of some of the variables may not be observed for all observations. For example, a variable may be observed for 75% of the study, but unobserved for the other 25%. Most of the techniques for analyzing missing data (multiple imputation, data augmentation, etc.) have been based on likelihood (Bayesian) methods.

The classical measurement error problem discussed above is one in which one set of variables, which we call U, is never observed, i.e., always missing. As such, the classical measurement error issue can be treated as a special kind of missing data problem, but with supplemental information in the form of surrogate, which we call W, and possibly a second measure, which we call T (Carroll et al. (2006)) . When we consider the measurement error problem, we are concerned with how the supplementary information is related to the unobserved covariate.

Because of the prescribed relationship between the two fields, and because missing data analysis has become increasingly parametric, it is reasonable to consider likelihood analysis of measurement error models (Carroll et al. (2006)). Likelihood methods require full statistical models for the distribution of U, sometime conditional on the observed covariates. Because these models describe the structure of U, they are called structural models. There are lots of concerns about the robustness of estimation and inference based on structural models for unobserved variables. Fuller (2009) discusses this issue briefly in the classical nonlinear regression problem,

and basically concludes that the results of structural modeling "may depend heavily on the form of the U distribution". In probit regression, Carroll et al. (1984) report that if one assumes that U is normally distributed, and it really follows a chi-squared distribution with one degree of freedom, then the effect on the likelihood estimate is markedly negative. Essentially all research workers in the measurement error field come to a common conclusion: likelihood methods can be considerably valuable, but the possible nonrobustness of inference due to model misspecification is a difficult problem.

The issue of model robustness is strictly limited to measurement error modelling. It has led to the rise of variety of semiparametric and nonparametric techniques. From this general point of view, functional modelling may be thought of as a group of semiparametric techniques. Functional modelling uses parametric models for the response, but makes no assumptions about the distribution of the unobserved covariate.

2.4.1 Maximum likelihood methods for measurement error

A likelihood analysis starts with determination of the joint distribution of Y, W given Z, as these are the observed variables. We first consider a simple problem where in Y, W and U are discrete random variables and there are no other covariates of Z. We know that

$$\begin{aligned} P(Y = y, W = w) &= \sum_u P(Y = y, W = w, U = u) \\ &= \sum_u P(Y = y | W = w, U = u) P(W = w, U = u). \end{aligned}$$

When W is a surrogate of U under the nondifferential measurement error assumption, it provides no additional information about Y when U is known, which means

Y is conditionally independent of W given the true covariate U, so that

$$P(Y = y, W = w) = \sum_u P(Y = y|U = u)P(W = w, U = u). \tag{2.3}$$

Therefore, we must specify a model for the joint distribution of W and U.

2.4.2 Error models

For additive and multiplicative error models, it is natural to specify the joint distribution of W and U in terms of the conditional distribution of W given U. Using the result from elementary probability that

$$P(W = w, U = u) = P(W = w|U = u)P(U = u). \tag{2.4}$$

Then (2.3) becomes

$$P(Y = y, W = w) = \sum_u P(Y = y|U = u)P(W = w|U = u)P(U = u). \tag{2.5}$$

Equation (2.5) has three components: (a) the main model of primary interest; (b) the error model for W given the true covariates U; (c) the distribution of the true covariates. Both (a) and (b) are expected. Almost all the methods for the measurement error process require a main model and an error model. However (c) is unexpected, in fact a bit disconcerting, because it requires a model for the distribution of the unobserved U. It is (c) that results in almost all the practical problems of implementation with maximum likelihood methods.

2.4.3 Berkson model

In the Berkson model, a univariate U is not observed but it is related to a univariate W by $U = W + e$, perhaps after a transformation. Usually, e is taken to be independent

of W and normally distributed with mean zero and variance σ^2, but more complex models are possible. When the Berkson model holds, we write

$$P(W = w, U = u) = P(U = u|W = w)P(W = w). \tag{2.6}$$

Then (2.3) becomes

$$P(Y = y, W = w) = \sum_u P(Y = y|U = u)P(U = u|W = w)P(W = w). \tag{2.7}$$

In equation (2.7), the third component is the distribution of W, and includes no information about the critical parameter of interest. Thus, we will divide both sides by $P(W = w)$ to get likelihoods conditional on W. In the general problem, we must specify the conditional density or mass function of U given W, which we denote by $f(u|w, \sigma)$. The likelihood function then becomes

$$f(y|w) = \int f(y|u)f(u|w)du. \tag{2.8}$$

In practice, summation or integral with respect to the distribution of U does not always yield an analytically closed form. Instead, we also can employ the Monte Carlo EM algorithm to solve this issue.

Chapter 3

Estimation for Zero-Inflated Beta-Binomial Regression Model with Missing Response Data

3.1 Introduction

The purpose of this chapter is to develop an estimation procedure for the parameters of the zero-inflated beta-binomial model with missing values. We consider all three missing data mechanisms. A weighted expectation maximization algorithm (Dempster et al. (1977)) is developed for the maximum likelihood (ML) estimation of the parameters involved. Extensive simulations are conducted to study the properties of the estimates using different measures, such as, average estimates (AE), relative bias (RB), variance (VAR), mean squared error (MSE) and coverage probability (CP) of estimates.

The zero-inflated beta-binomial model is introduced in Section 2. In this section

we also develop a procedure for the estimation of the parameters. Results of an extensive simulation study are reported in Section 3. Some illustrative examples are given in Section 4 and a discussion leading to some conclusions is given in Section 5.

3.2 The zero-inflated beta-binomial model and estimation procedure

3.2.1 The zero-inflated beta-binomial model

For a particular litter i, given m_i, the number of live foetuses in the litter, y_i, the number of foetuses affected, is a random variable having a beta-binomial distribution with parameter α and β, i.e,

$$f(y_i;\alpha,\beta)=\binom{m_i}{y_i}B(\alpha+y_i,m_i+\beta-y_i)/B(\alpha,\beta). \tag{3.1}$$

If $\pi=\frac{\alpha}{\alpha+\beta}$, and $\phi=\frac{1}{\alpha+\beta}$, we have

$$f(y_i;\alpha,\beta)=\binom{m_i}{y_i}\frac{\prod_{r=0}^{y_i-1}(\pi+r\phi)\prod_{r=0}^{m_i-y_i-1}(1-\pi+r\phi)}{\prod_{r=0}^{m_i-1}(1+r\phi)}, \tag{3.2}$$

with $E(Y_i)=m_i\pi$ and $Var(Y_i)=m_i\pi(1-\pi)\big[1+\frac{(m_i-1)\phi}{1+\phi}\big]$. We denote the beta-binomial distribution as $BB(\pi,\phi)$. As $\phi\to 0$ the $BB(\pi,\phi)$ tends to the binomial (π) distribution and for $\phi=0$ we have $Var(Y_i)=m_i\pi(1-\pi)$ and the $BB(\pi,\phi)$ becomes the binomial (π) distribution.

The zero-inflated beta binomial regression model (Deng and Paul (2005)) can be

written as

$$f(y_i|x_i;\pi,\phi,\omega) = \begin{cases} \omega + (1-\omega)\dfrac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi)}{\prod_{r=0}^{m_i-1}(1+r\phi)} & \text{if } y_i = 0, \\ (1-\omega)\binom{m_i}{y_i}\dfrac{\prod_{r=0}^{y_i-1}(\pi+r\phi)\prod_{r=0}^{m_i-y_i-1}(1-\pi+r\phi)}{\prod_{r=0}^{m_i-1}(1+r\phi)} & \text{if } y_i > 0, \end{cases} \tag{3.3}$$

with $E(Y_i) = (1-\omega)m_i\pi$, and $Var(Y_i) = (1-\omega)m_i\pi(1-\pi)\frac{1+m_i\phi}{1+\phi} + (1-\omega)\omega m_i^2\pi^2$, where ω is the zero-inflation parameter. We denote this distribution by $ZIBB(\pi,\phi,\omega)$. Inference regarding the parameters of the beta-binomial model and that of the zero-inflated beta-binomial model has been developed earlier (Dean (1992); Deng and Paul (2000)).

3.2.2 The estimation procedure

Suppose data from the $ZIBB(\pi,\phi,\omega)$ model for the i^{th} litter are (y_i, x_i), given the number m_i of litter size, $i = 1,\ldots,n$, y_i represents the response variable and x_i represents a $p\times 1$ vector of covariates with the regression parameter $\beta = (\beta_0,\beta_1,\ldots,\beta_p)$, such that $\pi_i = \exp(\sum_{j=0}^{p} X_{ij}\beta_j)/(1+\exp(\sum_{j=0}^{p} X_{ij}\beta_j))$. Here β_0 is the intercept parameter in which case $X_{i0} = 1$ for all i.

3.2.2.1 Estimation of ψ with no missing data

For complete data the log likelihood, apart from a constant, using the probability mass function given in equation (3.3), can be written as

$$\begin{aligned} l(\beta_j, \phi, \gamma | y_i) = \sum_{i=1}^{n} \Bigg[& -\log(1+\gamma) + \log\left[\gamma + \frac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1}(1+r\phi)}\right] I_{\{y_i=0\}} \\ & + \Big[\sum_{r=0}^{y_i-1}\log(\pi_i+r\phi) + \sum_{r=0}^{m_i-y_i-1}\log(1-\pi_i+r\phi) \\ & - \sum_{r=0}^{m_i-1}\log(1+r\phi)\Big] I_{\{y_i>0\}}\Bigg], \end{aligned} \tag{3.4}$$

where $\gamma = \omega/(1-\omega)$. Note, γ transforms the space of ω from $(0,1)$ onto $(0,\infty)$ which makes optimization of l easier (Deng and Paul (2005)). Let $\psi = (\beta, \phi, \gamma)$. Then the maximum likelihood estimates of the parameters ψ can be obtained by simultaneously solving the following estimating equations

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \Bigg[& \left[\frac{\Big(-\sum_{j=0}^{m_i-1}\prod_{r=0, r\neq j}^{m_i-1}(1+r\phi-\pi_i)\Big)}{\prod_{r=0}^{m_i-1}(1+r\phi)\left(\gamma + \frac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1}(1+r\phi)}\right)}\right] I_{\{y_i=0\}} \\ & + \left[\sum_{r=0}^{y_i-1}\frac{1}{\pi_i+r\phi} - \sum_{r=0}^{m_i-y_i-1}\frac{1}{1-\pi_i+r\phi}\right] I_{\{y_i>0\}}\Bigg] \frac{\partial \pi_i}{\partial \beta_j} = 0, \end{aligned}$$

$$\frac{\partial l}{\partial \phi} = \sum_{i=1}^{n} \Bigg[\Bigg[\frac{\Big(\sum_{j=0}^{m_i-1} j \prod_{r=0, r\neq j}^{m_i-1} (1+r\phi-\pi_i) \Big) \prod_{r=0}^{m_i-1} (1+r\phi)}{\Big(\prod_{r=0}^{m_i-1} (1+r\phi) \Big)^2 \Big(\gamma + \dfrac{\prod_{r=0}^{m_i-1} (1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1} (1+r\phi)} \Big)} - \frac{\Big(\sum_{j=0}^{m_i-1} j \prod_{r=0, r\neq j}^{m_i-1} (1+r\phi) \Big) \prod_{r=0}^{m_i-1} (1+r\phi-\pi_i)}{\Big(\prod_{r=0}^{m_i-1} (1+r\phi) \Big)^2 \Big(\gamma + \dfrac{\prod_{r=0}^{m_i-1} (1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1} (1+r\phi)} \Big)} \Bigg] I_{\{y_i=0\}} + \Bigg[\sum_{r=0}^{y_i-1} \frac{r}{\pi_i+r\phi} + \sum_{r=0}^{m_i-y_i-1} \frac{r}{1-\pi_i+r\phi} - \sum_{r=0}^{m_i-1} \frac{r}{1+r\phi} \Bigg] I_{\{y_i>0\}} \Bigg] = 0$$

and

$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{n} \Bigg[-(1+\gamma)^{-1} + \Bigg(\gamma + \frac{\prod_{r=0}^{m_i-1} (1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1} (1+r\phi)} \Bigg)^{-1} I_{\{y_i=0\}} \Bigg] = 0,$$

where $\dfrac{\partial \pi_i}{\partial \beta_j} = X_{ij} \exp(\sum_{j=0}^{p} X_{ij}\beta_j)/(1+exp(\sum_{j=0}^{p} X_{ij}\beta_j))^2$. Denote these estimates by $\hat{\psi}$. The observed information matrix of $\hat{\psi}$ is given by

$$H_0 = -\sum_{i=1}^{n} \frac{\partial^2}{\partial \psi \partial \psi'} l_i(\psi; y_i, x_i | \hat{\psi}). \tag{3.5}$$

The elements of this matrix are given in appendix 2. Of course, if it is convenient, these parameters can also be estimated by directly maximizing the log-likelihood function (3.4). However, in practice, through tests (Deng and Paul (2005)), if it is found the zero-inflation parameter is insignificant, then data analysis should be based on the beta-binomial model (3.2). The parameters β_j and ϕ can be estimated

by solving the estimating equations given in appendix 1. The elements of the observed information matrix corresponding to the model are also given in this appendix.

3.2.2.2 Estimation of the parameters with missing response

Under MCAR, the missingness is unrelated to the data. We can use the complete case (CC) analysis method which involves deletion of the cases that have missing values. The main advantage of this method is that it is easy to implement since we can use standard methods for complete data to compute the estimates. The disadvantage of the method is that we only use the cases that have complete information which may result in loss of efficiency of the estimates.

Note that in MAR missingness mechanisms are ignorable, which means that inference can proceed by analyzing the observed data only and without addressing the model for the missing data mechanism. As such MAR is a special case of MNAR for analyzing missing data. So, we first develop methods for MNAR in what follows and then obtain results for MAR by deleting the model for missing data mechanism.

As in Ibrahim et al. (2001) the complete data and missingness can be expressed as

$$y_i = \begin{cases} y_{o,i} & \text{if } y_i \text{ is observed,} \\ y_{m,i} & \text{if } y_i \text{ is missing.} \end{cases} \tag{3.6}$$

and

$$r_i = \begin{cases} 0 & \text{if } y_i \text{ is observed,} \\ 1 & \text{if } y_i \text{ is missing.} \end{cases} \tag{3.7}$$

We suppose the observed response and missing response have the same distribution $ZIBB(\pi, \phi, \omega)$ and missing data indicator r_i as follows

$$f(r_i|y_i, x_i; \alpha) = (p_i)^{r_i}(1-p_i)^{1-r_i}, \tag{3.8}$$

where $p_i = P(r_i = 1)$. To connect the distribution of r_i to covariates, logistic regression is employed,

$$log\big[\frac{P(r_i = 1)}{1 - P(r_i = 1)}\big] = Z_i^T \alpha, \tag{3.9}$$

where Z_i^T includes both missing data and observed data information, α is the vector of parameters of the missing data process. Let Y be the vector of responses, X be the covariate vector, Y_o be the vector of observed responses, Y_m be the vector of missing responses, and R be the vector of missing data indicators. Then, the full data density is given by

$$f(Y, R|X; \psi, \alpha) = f(Y|X; \psi) f(R|Y, X; \alpha) \tag{3.10}$$

where $\psi = (\beta, \phi, \gamma)$, and therefore the observed data density function can be written as

$$f(Y_o, R|X; \psi, \alpha) = \sum_{Y_m} f(Y|X; \psi) f(R|Y, X; \alpha). \tag{3.11}$$

Thus, the observed data log-likehood can be written as

$$l(\psi, \alpha|Y_o, R, X) = \log \sum_{Y_m} f(Y|X; \psi) f(R|Y, X; \alpha). \tag{3.12}$$

However, in practice, summation with respect to the distribution of Y_m is not always straight forward. An easier method is to use the EM algorithm of Dempster et al. (1977) which is developed below.

First, we write down the complete data log likelihood as

$$\begin{aligned} l(\beta, \phi, \gamma, \alpha|Y, R) &= \sum_{i=1}^{n} \Big[- \log(1 + \gamma) + \log[\gamma + f(0; \pi_i, \phi, \omega)] I_{\{y_i = 0\}} \\ &\quad + \log f(y_i; \pi_i, \phi, \omega) I_{\{y_i > 0\}} \Big] + \sum_{i=1}^{n} \Big[r_i Z_i^T \alpha - log(1 + e^{Z_i^T \alpha}) \Big]. \end{aligned} \tag{3.13}$$

The E-step provides the conditional expectation of the complete data log-likelihood with respect to the distribution of Y_m given the observed data and the current estimates of the parameters. Let s be an arbitrary number of iterations during maximization of the log-likelihood. Then given the observed data (Y_o, X, R) and current estimates of the parameters $\psi^{(s)}$ and $\alpha^{(s)}$, the conditional expectation of the complete data log-likelihood $l(\beta, \phi, \gamma, \alpha)$ for the i^{th} missing response in the $(s+1)^{th}$ iteration can be written as

$$\begin{aligned} Q_i(\psi, \alpha|\psi^{(s)}, \alpha^{(s)}) = & \; E\Big[l_i(\psi, \alpha; y_{o,i}, y_{m,i}, x_i, r_i|y_{o,i}, x_i, r_i; \psi^{(s)}, \alpha^{(s)})\Big] \\ = & \sum_{y_{m,i}=0}^{m_i} l_i(\psi, \alpha; y_{o,i}, y_{m,i}, x_i, r_i) f(y_{m,i}|y_{o,i}, x_i, r_i; \psi^{(s)}, \alpha^{(s)}). \end{aligned} \tag{3.14}$$

Suppose k of the n responses are observed and $n-k$ responses are missing. The responses are independent. Then, the E-step of the EM algorithm in the $(s+1)^{th}$ iteration is

$$\begin{aligned} Q(\psi, \alpha|\psi^{(s)}, \alpha^{(s)}) = & \sum_{i=1}^{k} l_i(\psi, \alpha; y_{o,i}, r_i, x_i) \\ & + \sum_{i=k+1}^{n} \sum_{y_{m,i}=0}^{m_i} l_i(\psi, \alpha; y_{m,i}, r_i, x_i) f(y_{m,i}|x_i, r_i; \psi^{(s)}, \alpha^{(s)}), \end{aligned} \tag{3.15}$$

where, using Bayes's theorem,

$$f(y_{m,i}|x_i, r_i; \psi^{(s)}, \alpha^{(s)}) = \frac{f(y_{m,i}|x_i; \psi^{(s)}) f(r_i|x_i, y_{m,i}; \alpha^{(s)})}{\sum_{y_{m,i}=0}^{m_i} f(y_{m,i}|x_i; \psi^{(s)}) f(r_i|x_i, y_{m,i}; \alpha^{(s)})}, \tag{3.16}$$

so that $Q(\psi, \alpha|\psi^{(s)}, \alpha^{(s)})$ can be expressed as

$$\begin{aligned} Q(\psi, \alpha|\psi^{(s)}, \alpha^{(s)}) = & \sum_{i=1}^{k} l_i(\psi, \alpha; y_{o,i}, r_i, x_i) \\ & + \sum_{i=k+1}^{n} \sum_{y_{m,i}=0}^{m_i} w_{iy_i}^{(s)} l_i(\psi, \alpha; y_{m,i}, r_i, x_i), \end{aligned} \tag{3.17}$$

where

$$w_{iy_i}^{(s)} = \frac{f(y_{m,i}|x_i;\psi^{(s)})f(r_i|x_i,y_{m,i};\alpha^{(s)})}{\sum_{y_{m,i}=0}^{m_i} f(y_{m,i}|x_i;\psi^{(s)})f(r_i|x_i,y_{m,i};\alpha^{(s)})}. \tag{3.18}$$

The M-step maximizes the function (3.13) with each log-likelihood for missing response being replaced by (m_i+1) weighted log-likelihood, where (m_i+1) is the number of distinct responses that missing observation i could have with different probabilities. If convergence is attained, then $\psi^{(s+1)}$ and $\alpha^{(s+1)}$ are the maximum likelihood estimates of the parameters ψ and α at the $(s+1)^{th}$ iteration. Denote these by $\hat{\psi}_1$ and $\hat{\alpha}$.

The variance-covariance matrix of the estimates of the parameters are obtained by inverting the observed information matrix at convergence (Efron and Hinkley (1978)), which is

$$\begin{aligned} H_1 = -Q''(\psi,\alpha|\psi^{(s)},\alpha^{(s)}) = & -\sum_{i=1}^{k} \frac{\partial^2}{\partial\psi\partial\psi'} l_i(\psi,\alpha;y_{o,i},x_i,r_i|\hat{\psi}_1,\hat{\alpha}) \\ & -\sum_{i=k+1}^{n}\sum_{y_{m,i}=0}^{m_i} w_{iy_i}^{(s)} \frac{\partial^2}{\partial\psi\partial\psi'} l_i(\psi,\alpha;y_{m,i},x_i,r_i|\hat{\psi}_1,\hat{\alpha}). \end{aligned} \tag{3.19}$$

Expressions for the elements of H_1 regarding to estimates $\hat{\psi}_1$ are given in Appendix 3 by replacing the parameters β, ϕ, and γ with $\hat{\psi}_1(\ \hat{\beta},\ \hat{\phi}$, and $\hat{\gamma})$.

In case of MAR the corresponding results for the estimates of ψ, after deleting the model for the missing data mechanism, are obtained as follows:

The E-step: Given the observed data (X) and current estimates of the parameters $\psi^{(s)}$, the conditional expectation of the complete data log-likelihood $l(\beta,\phi,\gamma)$ for the i^{th} missing response in the $(s+1)^{th}$ iteration is

$$\begin{aligned} Q_i(\psi|\psi^{(s)}) = & \ E\Big[l_i(\psi;y_{o,i}y_{m,i},x_i|x_i;\psi^{(s)})\Big] \\ = & \sum_{y_{m,i}=0}^{m_i} l_i(\psi;y_{o,i},y_{m,i},x_i) f(y_{m,i}|x_i;\psi^{(s)}), \end{aligned} \tag{3.20}$$

which for all the observations is

$$Q(\psi|\psi^{(s)}) = \sum_{i=1}^{k} l_i(\psi; y_{o,i}, x_i) + \sum_{i=k+1}^{n} \sum_{y_{m,i}=0}^{m_i} w_{iy_i}^{(s)} l_i(\psi; y_{m,i}, x_i), \tag{3.21}$$

where

$$w_{iy_i}^{(s)} = f(y_{m,i}|x_i, \psi^{(s)}). \tag{3.22}$$

The M-step maximizes the function (3.4) with each missing response being replaced by $(m_i + 1)$ weighted observations, where $(m_i + 1)$ is the number of distinct responses that missing observation i could have with different probabilities. If convergence is attained, then $\psi^{(s+1)}$ is the maximum likelihood estimate of the parameters ψ at the $(s+1)^{th}$ iteration. Denote this by $\hat{\psi}_2$.

The observed information matrix of the estimates $\hat{\psi}$ is

$$\begin{aligned} H_2 = -Q''(\psi|\psi^{(s)}) = & -\sum_{i=1}^{k} \frac{\partial^2}{\partial\psi\partial\psi'} l_i(\psi; y_{o,i}, x_i|\hat{\psi}_2) \\ & - \sum_{i=k+1}^{n} \sum_{y_{m,i}=0}^{m_i} w_{iy_i}^{(s)} \frac{\partial^2}{\partial\psi\partial\psi'} l_i(\psi; y_{m,i}, x_i|\hat{\psi}_2) \end{aligned} \tag{3.23}$$

Expressions for the elements of H_2 are given in Appendix 3 by replacing the parameters β, ϕ, and γ with $\hat{\psi}_2$($\hat{\beta}$, $\hat{\phi}$, and $\hat{\gamma}$).

3.3 Simulation study

A simulation study was conducted to investigate the properties of the estimates in terms of average estimates (AE), relative bias (RB), variance (VAR), mean squared error (MSE) and coverage probability (CP) of estimates. The AE, RB, SE, MSE and CP, for example of $\hat{\pi}$, are obtained as: AE $(\hat{\pi}) = \frac{1}{N}\sum_{q=1}^{N} \hat{\pi}_q$, RB($\hat{\pi}$)= (AE - π)/π, VAR($\hat{\pi}$) = $\frac{1}{N}\sum_{q=1}^{N} \widehat{var}(\hat{\pi}_q)$, where $\widehat{var}(\hat{\pi}_q)$ was obtained from the observed

information matrix given in (3.19) or (3.23), $\text{MSE}(\hat{\pi}) = \frac{1}{N}\sum_{q=1}^{N}(\hat{\pi}_q - \pi)^2$, and $\text{CP}(\hat{\pi}) = \frac{1}{N}\sum_{q=1}^{N} I(\hat{\pi}_q - Z_{\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\pi}_q)} < \pi < \hat{\pi}_q + Z_{\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\pi}_q)})$, where N is the number of samples we simulated.

We use data under four scenarios: (i) data are observed completely, (ii) some responses are missing completely at random (MCAR), (iii) some responses are missing at random (MAR), (iv) some responses are missing not at random (MNAR).

Two sets of simulations are conducted. The first is with no covariate and the second is with a one covariate.

In the case in which there is no covariate, response data are generated from the zero-inflated beta binomial model (3.3) with $m_i = 10$, $\pi = 0.8$, $\phi = 0.2$ and $\omega = 0.2$. The missing data indicator r_i is generated independently by the following model

$$logit(P(r_i = 1)) = \alpha_0 + \alpha_1 y_i. \tag{3.24}$$

We set $\alpha_0 = (-3, -2.2, -1.1)$ which produces about $5\%, 10\%, 25\%$ missing observations at the baseline. The baseline missing rate is $P(r_i = 1)) = exp(\alpha_0)/(1+exp(\alpha_0))$. The parameter α_1 is set to 0 and 0.1 to indicate different missing data mechanisms MCAR(MAR) and MNAR respectively (Ibrahim and Lipsitz (1996)).

Note that when there is no covariate, we only have response data y_i. Thus from the missing data indicator model (3.24) we see that the missing data mechanism is unrelated to the data if $\alpha_1 = 0$ indicating that the missingness is MCAR. However, if $\alpha_1 \neq 0$, the missing data mechanism depends on the unobserved response y_i when y_i is missing, which results in nonignorable missing data mechanism MNAR.

For the case with one covariate we take $\pi_i = \exp(\beta_0 + \beta_1 x_i)/(1 + \exp(\beta_0 + \beta_1 x_i))$ with $\beta_0 = -1$, $\beta_1 = 1$. Note that β_0 is the intercept parameter. The regression variable x_i was generated from $N(1, 1)$. For the missing data process, we consider the

logistic model

$$logit(P(r_i = 1)) = \alpha_0 + \alpha_1 x_i + \alpha_2 y_i \tag{3.25}$$

from which missing data indicators r_i's are independently generated. The value of α_0 is set the same as in the case with no covariate. The values of (α_1, α_2) are set as $(0, 0), (0.1, 0), (-0.1, 0.1)$ to indicate missing data mechanism MCAR, MAR and MNAR respectively (Ibrahim and Lipsitz (1996)).

Here also note from model (3.25) that, when $\alpha_1 = 0$ and $\alpha_2 = 0$, the missing data do not depend on either the observed covariate x_i or the missing response y_i, which results in MCAR. When $\alpha_1 \neq 0$ and $\alpha_2 = 0$, the missingness only depends on the observed covariate x_i resulting in MAR. When $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$, the missingness depends on the missing response y_i, in addition to the observed covariate x_i indicating that we have MNAR. Here, in order to control the missing rate close to the baseline missing rate, we set small values for α_1 and α_2.

For empirical coverage probability we take nominal level $\alpha = 0.05$.

When there is no covariate, simulation results for complete data, data under (MCAR and MAR) are given in Table 3.1 to Table 3.3. Simulation results under MNAR are given in Table 3.4 to Table 3.6. When there is one covariate, the corresponding results for complete data (also MCAR), MAR and MNAR are given in Table 3.7 to Table 3.15 respectively. In each case data were analyzed by the CC, EM-MCAR (MAR) and the EM-MNAR method.

We first discuss the results in Tables 3.1 to Table 3.6 for the situations in which there is no covariate.

Results in Table 3.1 to Table 3.3 for complete data indicate that all the parameters are well estimated irrespective of the sample sizes and at $n = 100$ the result shows almost no estimation error. However, the coverage probability falls short of the

nominal coverage of 95%.

The parameters π and ω are well estimated irrespective of percentage missing and sample size. All of AE, RB, VAR, and MSE show good behavior for all sample sizes and percentage missing does not seem to have any effect on these. However, the coverage probability decreases somewhat as percentage missing increases, although never falls below 92%.

The parameter ϕ shows relatively high RB (as high as 8%) and slightly higher (3%) VAR and MSE and shorter coverage probability for smaller n (n=30). As the sample size increases (n=100) all other indices show good properties, although still conservative in terms of coverage probability, particularly as percentage missing increases (or sample size decreases). Its CP ranges from .88 to .93. Note that difference in the coverage probability between 25% missing and that for 50% missing is very small (0.89 for 25% and 0.88 for 50%).

Note that when there is no covariate, increase in percentage missing under MCAR or MAR has the same effect as reducing the sample size. So, under MCAR and MAR these results (CC method) should be very similar to those if the EM method is applied to replace the missing observations. To confirm this we included the results using the EM method in Table 3.1 to Table 3.3 (and in subsequent tables). The simulation results obtained by analyzing with EM-MNAR are very similar with those analyzed under EM-MAR(MCAR). After round off (up) to three decimals, the results are same, which means the EM MNAR works well under MCAR.

In Table 3.4 to Table 3.6 when there is no covariate, the results show that when the data are simulated under MNAR but analyzed by the CC method (complete case analysis method) or the EM-MAR(MACR), it yields considerably larger AE, RB, SE and MSE and lower coverage probability, even for large sample size. The parameter

π shows underestimation, whereas, the other two parameters show overestimation.

The EM-MNAR method, however, shows excellent performance in terms of all the measures for all three parameter estimates, except that the coverage probability for the parameter ϕ is shorter (ranges from .87 to .91) in comparison to that from complete data. However, these coverage probabilities are much closer to the nominal coverage probability than those using the CC method. All parameters are well estimated even at 25% baseline missing.

We next discuss the results in Table 3.7 to Table 3.15 for the situations in which there is one covariate.

Results in Table 3.7 to Table 3.9 show that the parameters β_0, β_1 and ω are well estimated irrespective of percentage missing and sample size. All of AE, RB, VAR, and MSE show good behavior for all sample sizes and percentage missing does not seem to have any effect on these. These properties are very similar to those of π and ω given in Table 3.1 to Table 3.3 where there was no covariate. However, the coverage probability decreases further than those given in Table 1 as percentage missing increases, although never falls below 90%.

The parameter ϕ shows high RB (as high as 23%) for small sample size($n = 30$) and large missing percentage (50%). However, as the sample size increases ($n = 100$) RB decreases to 5%. The behavior of VAR and that of MSE are similar to those in Table 3.1 to Table 3.3, namely, that these are slightly higher than those for complete data. As the sample size increases (n=100) all other indices show good properties, although still conservative in terms of coverage probability, particularly as percentage missing increases (or sample size decreases). Its CP ranges from .81 to .93. The difference in the coverage probability between 25% missing and that for 50% missing is small (0.85 for 25% and 0.81 for 50%).

Note that the simulation results in Table 3.1 to Table 3.3 and Table 3.7 to Table 3.9 show that for all other parameters except ϕ, the properties of the estimates for 50% missing are similar to those for 25% missing. For ϕ, only for coverage probability, some difference is shown. This seems to be the pattern. So, in all other tables we do not include simulation results for 50% missing.

Similarly, estimates of all the parameters, under MAR and MNAR, results of which are given in Table 3.10 and Table 3.15, show similar behavior as those in Table 3.4 to Table 3.6 except that it now requires much larger sample sizes.

In summary, Under MCAR, both the EM methods (EM-MCAR(MAR) and EM-MNAR) and the CC method work well. However, the EM methods are more time consuming compared to the CC method. The EM MAR(MCAR) method performs well for missing data under MCAR and MAR, but produces bias under MNAR. The EM-MNAR performs well under all three missing data mechanisms.

3.4 An Example: Analysis of a mutagenic data set

In this section we analyze a set of mutagenic data. The data obtained from Lüning et al. (1966) involved groups of male mice originating from an inbred CBA strain mated with groups of female mice originating from same inbred CBA strain. The experiment was conducted in three groups in which male mice were given 0 R, 300 R and 600 R respectively and then were mated within the first 7 days after irradiation.

The data are given in Table 3.16, and grouped according to the number of implants and the number of dead foetuses. We are interested in the dosage effect on the death rate of the foetuses. The outcome variable is the number of dead foetuses in the litter. The independent variable is the dosage.

Our purpose here is to illustrate analysis of zero-inflated beta-binomial data with missing values in the response variable. However, we first analyze the complete data using the zero-inflated beta-binomial model (3.3) with $\pi_i = \exp(\beta_0 + \beta_1 x_i)/(1 + \exp(\beta_0 + \beta_1 x_i))$, $x_i = treatment_i = 0, 300, 600$, where π_i is the proportion of dead implants, β_0 represents the intercept parameter and β_1 represents the regression parameter (treatment effect). Since the dosages x_i are far apart we standardize as $z_i = (x_i - \bar{x})/s$, where $\bar{x}$ and s are mean and standard deviation of the x_i values.

The model then for the zero-inflated beta-binomial proportion becomes $\pi_i = \exp(\beta_0+\beta_1 z_i)/(1+\exp(\beta_0+\beta_1 z_i))$. The maximum likelihood estimate (mle) of β_0, β_1, ϕ and ω for the mutagenic data in Table 3.16 are -1.314, 0.702, 0.026, and $1.206{*}10^{-6}$ respectively. It seems that the zero-inflation parameter does not contribute much to the model. Further evidence of such insignificance of ω has been found by testing $H_0 : \omega = 0$ using the score test statistic Z_8 given in Deng and Paul (2005). This statistic has an asymptotic chi-square distribution with one degree of freedom and for our data $Z_8 = 1.104$ confirming that the zero-inflation parameter is not significant. We further test whether the over-dispersion parameter ϕ is significant by using the score test statistic Z_7 of Deng and Paul (2005), which also has an asymptotic chi-square distribution with one degree of freedom. Its value for the data in Table 3.16 is $Z_7 = 18.589$ with a p-value of $9.28 * 10^{-4}$ indicating significance at 5% level.

Two sets of analyses with missing responses are performed. First, note that the BB model fits the data set in Table 3.16, but does not contain any missing values. However, in practice, in Toxicology and mutagenic studies, missingness can occur in addition to the data being over-dispersed. So, to illustrate our method of analyzing mutagenic or toxicological data in the form of proportions that follow the BB model, but contain missing responses we generate missingness using the model

(3.25). Estimates of the parameters β_0, β_1, and ϕ and their variances using the CC, EM-MAR (MCAR), and the EM-MNAR methods are given in Tables 3.17(a), 3.17(b) and 3.17(c) for MCAR, MAR and MNAR respectively.

Results in Table 3.17(a) indicate that percentage missing has a some small effect on the mle and the estimate of its variance. That is, the mle and the estimate of its variance remain relatively stable even at 25% missing.

Results in Table 3.17(b) indicate that percentage missing has some effect on the mle and the estimate of its variance under MAR when analyzed under the CC method. However, these results remain almost unaffected when the missing data are replaced by their estimates using the EM-MAR(MCAR) and EM-MNAR method. Results in Table 3.17(c) indicate that percentage missing has some effect on the mle and the estimate of its variance under MNAR when analyzed under the CC method and EM-MAR(MCAR) method. However, these results remain almost unaffected when the missing data are replaced by their estimates using the EM-MNAR method.

In the second set of analysis, we first generate a new data set from the zero-inflated over-dispersed beta-binomial model (3.3) using the implantation sizes and treatments as in Table 6 and the values $\beta_0 = -1.314$, $\beta_1 = 0.702$, and $\phi = 0.026$, obtained as mles from the data in Table 3.16. The zero-inflation parameter ω was set as $\omega = 0.03$. These data are given in Table 3.18. We then test whether both the zero-inflation and over-dispersion parameters are significant in these data. The mles of β_0, β_1, ϕ and ω for these data are -1.307, 0.660, 0.020, and 0.033 respectively. The values of the score test statistics for testing for over-dispersion and for zero-inflation are $Z_7 = 12.715$ and $Z_8 = 10.601$ respectively with p-values of 0.0103 and 0.0241 indicating significance of the over-dispersion and zero-inflation parameters at 5% level of significance. So, we proceed with model (3.3) to analyze these data and study the impact of missing

data. As earlier, for incomplete data, we generate missingness using the model (3.25). Estimates of the parameters for the new data set and that with missing data under MCAR, MAR and MNAR and analyzed using the CC, EM-MAR (MCAR), and the EM-MNAR methods are given in Tables 3.19 (a), 3.19 (b) and 3.19 (c) for respectively.

Results in Table 3.19 (a) under MCAR indicate that, as in Table 3.17(a), percentage missing has a small effect on the mle's and estimates of their variances. Results in Table 3.19 (b) under MAR, show that percentage missing could have significant effect on the mles and estimates of their variances of all the parameters when we use the complete case (CC) method. However, as in Tables 3.17(b), these results remain almost unaffected when the missing data are replaced by their estimates using the EM-MAR(MCAR) and EM-MNAR method. Results in Table 3.19 (c) under MNAR, show that percentage missing could have significant effect on the mles and estimates of their variances of all the parameters when we use the complete case (CC) method and EM-MAR(MCAR) method. However, as in Tables 3.17(c), these results remain almost unaffected when the missing data are replaced by their estimates using the EM-MNAR method.

A question may arise why do we not analyze the data using a zero-inflated negative binomial model or a zero-inflated generalized Poisson model as there is over-dispersion in the data ($\bar{y}$=1.51 and $s^2 = 1.74$ when we ignore the binomial denominators). The drawback of these Poisson related models is that the data ignore the binomial denominators and the conditions, p is small and n large (for example, $n \geq 20$ and $p \leq 0.05$) (Hogg et al. (1977)) are generally violated to approximate a binomial (n, p) distribution by a Poisson (np) distribution. To check this point, we analyzed these data by using a negative binomial model with over-dispersion parameter c, a zero-inflated Poisson model with zero-inflation parameter ω and a zero-inflated negative

binomial model with over-dispersion parameter c and zero-inflation parameter ω. The maximum likelihood estimates of these parameters in the three models are $c = 2.39 * 10^{-8}, \omega = 1.12 * 10^{-5}$, and $c = 1.19 * 10^{-7}, \omega = 3.36 * 10^{-7}$ respectively. The estimates of the over-dispersion and/or the zero-inflation parameters are very close to zero. This shows that when we analyze the over-dispersed and or zero-inflated binomial data by an over-dispersed and or a zero-inflated Poisson model, the analysis may not capture all important features of the data.

3.5 Discussion

We develop estimation procedure for the parameters of a zero-inflated beta-binomial model in presence of missing responses. We apply a weighted expectation maximization algorithm for the maximum likelihood estimation of the parameters. Although missing data methodologies have been developed extensively in the literature, the current development for the estimation of the parameters of ZIBB in presence of missing responses is new. For completeness we also discuss, in Section 2 and 4, how to deal with the missing data under a beta-binomial model.

An extensive simulation study and analysis of some illustrative data sets are performed. In both simulations and data analyses, complete data and data with missing values under MCAR, MAR and MNAR with or without covariates are considered.

The general findings through simulations and data analyses are:

(a) Data without covariates: for complete data and under MCAR and MAR, all the parameters are well estimated irrespective of the sample sizes and percentage missing. All of the AE, RB, VAR, and MSE show good behavior. However, all the parameter estimates show shorter coverage probability, especially for ϕ, whose

coverage probability ranges from 0.91 to 0.93. Under MNAR, the CC method for all the parameters yields considerably larger AE, RB, SE and MSE and lower coverage probability, even for large sample size. The EM method shows excellent performance in terms of all the measures for all three parameter estimates, except that the coverage probability for the parameter ϕ is shorter (ranges from .87 to .91) in comparison to that from complete data. However, these coverage probabilities are much closer to the nominal coverage probability than those using the CC method. All parameters are well estimated even at 25% baseline missing.

(b) Data with one covariate: Results for complete data are almost the same as those with no covariate except that to see such good behavior much larger sample sizes are required. Similarly, estimates of all the parameters, under MAR and MNAR, show similar behavior as those with no covariates except now require much larger sample sizes.

Analyses of two data sets, one that fits a beta-binomial model and the other that fits a zero-inflated beta-binomial model show similar findings.

Table 3.1: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters for complete data and data under MCAR (also MAR) with no covariate, data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 30$).*

		Complete data			5% missing			10% missing			25% missing			50% missing		
Method	Quantity	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$
n=30	AE	0.797	0.208	0.201	0.797	0.208	0.201	0.796	0.209	0.201	0.796	0.209	0.199	0.793	0.216	0.202
	RB	-0.004	0.037	0.004	-0.004	0.038	0.004	-0.004	0.045	0.005	-0.004	0.046	-0.003	-0.008	0.079	0.011
CC	VAR	0.002	0.013	0.005	0.002	0.013	0.005	0.002	0.014	0.006	0.002	0.018	0.007	0.004	0.032	0.010
	MSE	0.002	0.013	0.005	0.002	0.013	0.005	0.002	0.014	0.006	0.002	0.018	0.007	0.004	0.033	0.010
	CP	0.936	0.913	0.962	0.935	0.907	0.945	0.933	0.903	0.933	0.923	0.892	0.931	0.929	0.880	0.929
	AE	0.797	0.208	0.201	0.797	0.208	0.201	0.796	0.209	0.201	0.796	0.209	0.199	0.793	0.216	0.202
	RB	-0.004	0.037	0.004	-0.004	0.038	0.004	-0.004	0.045	0.005	-0.004	0.046	-0.003	-0.008	0.079	0.011
EM-MCAR(MAR)	VAR	0.002	0.013	0.005	0.002	0.012	0.005	0.002	0.013	0.005	0.002	0.016	0.005	0.004	0.028	0.007
	MSE	0.002	0.013	0.005	0.002	0.012	0.005	0.002	0.013	0.005	0.002	0.016	0.005	0.004	0.029	0.008
	CP	0.936	0.913	0.962	0.935	0.902	0.945	0.933	0.901	0.931	0.923	0.891	0.932	0.930	0.881	0.913
	AE	0.797	0.208	0.201	0.797	0.208	0.201	0.796	0.209	0.201	0.796	0.209	0.199	0.793	0.216	0.202
	RB	-0.004	0.037	0.004	-0.004	0.038	0.004	-0.004	0.045	0.005	-0.004	0.046	-0.003	-0.008	0.079	0.011
EM-MNAR	VAR	0.002	0.013	0.005	0.002	0.012	0.005	0.002	0.013	0.005	0.002	0.016	0.005	0.004	0.028	0.007
	MSE	0.002	0.013	0.005	0.002	0.012	0.005	0.002	0.013	0.005	0.002	0.016	0.005	0.004	0.029	0.008
	CP	0.936	0.913	0.962	0.935	0.902	0.945	0.933	0.901	0.931	0.923	0.891	0.932	0.930	0.881	0.913

CC, complete case analysis.

Table 3.2: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters for complete data and data under MCAR (also MAR) with no covariate, data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 50$).*

		Complete data			5% missing			10% missing			25% missing			50% missing		
Method	Quantity	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$
n=50	AE	0.799	0.201	0.201	0.799	0.201	0.201	0.799	0.200	0.202	0.799	0.200	0.201	0.798	0.205	0.199
	RB	-0.001	0.005	0.006	-0.001	0.006	0.007	-0.001	0.001	0.010	-0.001	-0.001	0.004	-0.002	0.023	-0.003
CC	VAR	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.008	0.004	0.001	0.009	0.004	0.002	0.016	0.006
	MSE	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.008	0.004	0.001	0.009	0.004	0.002	0.016	0.006
	CP	0.936	0.914	0.943	0.935	0.913	0.927	0.930	0.907	0.931	0.932	0.894	0.927	0.926	0.885	0.918
	AE	0.799	0.201	0.201	0.799	0.201	0.201	0.799	0.200	0.202	0.799	0.200	0.201	0.798	0.205	0.199
	RB	-0.001	0.005	0.006	-0.001	0.006	0.007	-0.001	0.001	0.010	-0.001	-0.001	0.004	-0.002	0.023	-0.003
EM-MCAR(MAR)	VAR	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.007	0.003	0.002	0.014	0.004
	MSE	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.007	0.003	0.002	0.014	0.004
	CP	0.936	0.914	0.943	0.935	0.913	0.927	0.930	0.905	0.930	0.932	0.895	0.924	0.926	0.886	0.915
	AE	0.799	0.201	0.201	0.799	0.201	0.201	0.799	0.200	0.202	0.799	0.200	0.201	0.798	0.205	0.199
	RB	-0.001	0.005	0.006	-0.001	0.006	0.007	-0.001	0.001	0.010	-0.001	-0.001	0.004	-0.002	0.023	-0.003
EM-MNAR	VAR	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.007	0.003	0.002	0.014	0.004
	MSE	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.007	0.003	0.001	0.007	0.003	0.002	0.014	0.004
	CP	0.936	0.914	0.943	0.935	0.913	0.927	0.930	0.905	0.930	0.932	0.895	0.924	0.926	0.886	0.915

CC, complete case analysis.

Table 3.3: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters for complete data and data under MCAR (also MAR) with no covariate, data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 100$).*

		Complete data			5% missing			10% missing			25% missing			50% missing		
Method	Quantity	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$
n=100	AE	0.800	0.200	0.200	0.800	0.200	0.200	0.800	0.200	0.200	0.800	0.201	0.200	0.800	0.201	0.201
CC	RB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.004	0.000	0.000	0.003	0.007
	VAR	0.001	0.003	0.002	0.001	0.004	0.002	0.001	0.004	0.002	0.001	0.005	0.002	0.001	0.007	0.003
	MSE	0.001	0.003	0.002	0.001	0.004	0.002	0.001	0.004	0.002	0.001	0.005	0.002	0.001	0.007	0.003
	CP	0.943	0.926	0.936	0.942	0.925	0.936	0.941	0.924	0.937	0.941	0.923	0.935	0.935	0.909	0.931
EM-MCAR(MAR)	AE	0.800	0.200	0.200	0.800	0.200	0.200	0.800	0.200	0.200	0.800	0.201	0.200	0.800	0.201	0.201
	RB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.004	0.000	0.000	0.003	0.007
	VAR	0.001	0.003	0.002	0.001	0.003	0.002	0.001	0.003	0.002	0.001	0.004	0.002	0.001	0.006	0.003
	MSE	0.001	0.003	0.002	0.001	0.003	0.002	0.001	0.003	0.002	0.001	0.004	0.002	0.001	0.006	0.003
	CP	0.943	0.926	0.936	0.942	0.925	0.936	0.942	0.922	0.936	0.941	0.920	0.936	0.936	0.905	0.930
EM-MNAR	AE	0.800	0.200	0.200	0.800	0.200	0.200	0.800	0.200	0.200	0.800	0.201	0.200	0.800	0.201	0.201
	RB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.004	0.000	0.000	0.003	0.007
	VAR	0.001	0.003	0.002	0.001	0.003	0.002	0.001	0.003	0.002	0.001	0.004	0.002	0.001	0.006	0.003
	MSE	0.001	0.003	0.002	0.001	0.003	0.002	0.001	0.003	0.002	0.001	0.004	0.002	0.001	0.006	0.003
	CP	0.943	0.926	0.936	0.942	0.925	0.936	0.942	0.922	0.936	0.941	0.920	0.936	0.936	0.905	0.930

CC, complete case analysis.

Table 3.4: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under MNAR with no covariate, data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 30$).*

		Complete data			5% missing			10% missing			25% missing		
Method	Quantity	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$
n=30	AE	0.797	0.208	0.201	0.793	0.207	0.210	0.791	0.210	0.218	0.779	0.216	0.244
	RB	-0.004	0.037	0.004	-0.009	0.035	0.050	-0.011	0.050	0.090	-0.026	0.080	0.220
CC	VAR	0.002	0.013	0.005	0.002	0.014	0.006	0.002	0.016	0.007	0.003	0.025	0.010
	MSE	0.002	0.013	0.005	0.002	0.014	0.006	0.002	0.016	0.007	0.004	0.025	0.012
	CP	0.936	0.913	0.962	0.915	0.879	0.946	0.905	0.879	0.919	0.850	0.808	0.876
	AE	0.797	0.208	0.201	0.793	0.207	0.209	0.791	0.209	0.219	0.779	0.215	0.244
	RB	-0.004	0.037	0.004	-0.009	0.035	0.045	-0.011	0.045	0.095	-0.026	0.075	0.220
EM-MCAR(MAR)	VAR	0.002	0.013	0.005	0.002	0.012	0.005	0.002	0.013	0.005	0.002	0.014	0.006
	MSE	0.002	0.013	0.005	0.002	0.012	0.005	0.002	0.013	0.006	0.002	0.014	0.008
	CP	0.936	0.913	0.962	0.914	0.882	0.946	0.889	0.876	0.921	0.849	0.809	0.888
	AE	0.797	0.208	0.201	0.797	0.202	0.204	0.795	0.204	0.205	0.795	0.205	0.206
	RB	-0.004	0.037	0.004	-0.004	0.010	0.020	-0.006	0.020	0.025	-0.006	0.025	0.030
EM-MNAR	VAR	0.002	0.013	0.005	0.002	0.012	0.005	0.002	0.012	0.005	0.002	0.012	0.005
	MSE	0.002	0.013	0.005	0.002	0.012	0.005	0.002	0.012	0.005	0.002	0.012	0.005
	CP	0.936	0.913	0.962	0.940	0.888	0.946	0.930	0.905	0.927	0.931	0.884	0.939

CC, complete case analysis.

Table 3.5: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under MNAR with no covariate, data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 50$).*

		Complete data			5% missing			10% missing			25% missing		
Method	Quantity	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$
n=50	AE	0.799	0.201	0.201	0.794	0.206	0.209	0.792	0.208	0.216	0.782	0.212	0.236
	RB	-0.001	0.005	0.006	-0.008	0.030	0.045	-0.010	0.040	0.080	-0.023	0.060	0.180
CC	VAR	0.001	0.007	0.003	0.002	0.013	0.006	0.002	0.015	0.007	0.003	0.024	0.010
	MSE	0.001	0.007	0.003	0.002	0.013	0.006	0.002	0.015	0.007	0.004	0.024	0.011
	CP	0.936	0.914	0.943	0.926	0.883	0.931	0.903	0.856	0.915	0.835	0.810	0.871
	AE	0.799	0.201	0.201	0.795	0.202	0.209	0.792	0.207	0.212	0.782	0.212	0.236
	RB	-0.001	0.005	0.006	0.042	0.104	0.072	0.042	0.111	0.071	0.047	0.113	0.078
EM-MCAR(MAR)	VAR	0.001	0.007	0.003	0.002	0.011	0.005	0.002	0.012	0.005	0.002	0.013	0.006
	MSE	0.001	0.007	0.003	0.002	0.011	0.005	0.002	0.012	0.005	0.003	0.013	0.007
	CP	0.936	0.914	0.943	0.925	0.881	0.930	0.913	0.860	0.900	0.834	0.812	0.881
	AE	0.799	0.201	0.201	0.798	0.202	0.201	0.796	0.203	0.203	0.798	0.204	0.204
	RB	-0.001	0.005	0.006	0.041	0.106	0.071	0.041	0.107	0.071	0.041	0.113	0.071
EM-MNAR	VAR	0.001	0.007	0.003	0.002	0.011	0.005	0.002	0.011	0.005	0.002	0.013	0.005
	MSE	0.001	0.007	0.003	0.002	0.011	0.005	0.002	0.011	0.005	0.002	0.013	0.005
	CP	0.936	0.914	0.943	0.938	0.895	0.929	0.931	0.895	0.931	0.927	0.874	0.946

CC, complete case analysis.

Table 3.6: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under MNAR with no covariate, data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 100$).*

		Complete data			5% missing			10% missing			25% missing		
Method	Quantity	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$	$\pi = 0.8$	$\phi = 0.2$	$\omega = 0.2$
n=100	AE	0.800	0.200	0.200	0.795	0.204	0.207	0.794	0.207	0.210	0.785	0.210	0.220
	RB	0.000	0.000	0.000	-0.006	0.020	0.035	-0.008	0.035	0.050	-0.019	0.050	0.100
CC	VAR	0.001	0.003	0.002	0.002	0.014	0.006	0.002	0.013	0.007	0.003	0.020	0.010
	MSE	0.001	0.003	0.002	0.002	0.014	0.006	0.002	0.013	0.007	0.004	0.020	0.010
	CP	0.943	0.926	0.936	0.913	0.885	0.936	0.892	0.870	0.927	0.841	0.831	0.850
	AE	0.800	0.200	0.200	0.796	0.203	0.208	0.795	0.206	0.210	0.784	0.209	0.219
	RB	0.000	0.000	0.000	-0.005	0.015	0.040	-0.006	0.030	0.050	-0.020	0.045	0.095
EM-MCAR(MAR)	VAR	0.001	0.003	0.002	0.002	0.012	0.005	0.002	0.012	0.005	0.003	0.013	0.008
	MSE	0.001	0.003	0.002	0.002	0.014	0.006	0.002	0.013	0.005	0.003	0.023	0.008
	CP	0.943	0.926	0.936	0.912	0.885	0.937	0.891	0.872	0.928	0.842	0.830	0.852
	AE	0.800	0.200	0.200	0.799	0.201	0.201	0.798	0.202	0.202	0.797	0.203	0.202
	RB	0.000	0.000	0.000	-0.001	0.005	0.005	-0.003	0.010	0.010	-0.004	0.015	0.010
EM-MNAR	VAR	0.001	0.003	0.002	0.002	0.012	0.005	0.002	0.012	0.005	0.002	0.013	0.005
	MSE	0.001	0.003	0.002	0.002	0.012	0.005	0.002	0.013	0.005	0.002	0.013	0.005
	CP	0.943	0.926	0.936	0.930	0.895	0.935	0.929	0.900	0.939	0.922	0.899	0.931

CC, complete case analysis.

Table 3.7: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters for complete data and data under MCAR with one covariate, data simulated from BB(β_0, β_1, ϕ, ω), based on 1000 simulation runs ($n = 30$).*

		Complete data				5% missing				10% missing				25% missing				50% missing			
Method	Quantity	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$
n=30	AE	-1.026	1.015	0.179	0.207	-1.025	1.018	0.177	0.207	-1.028	1.016	0.174	0.206	-1.028	1.024	0.170	0.208	-1.030	1.040	0.154	0.213
CC	RB	0.026	0.015	-0.105	0.035	0.025	0.018	-0.113	0.034	0.028	0.016	-0.128	0.028	0.028	0.024	-0.150	0.039	0.030	0.040	-0.229	0.064
	VAR	0.154	0.082	0.010	0.008	0.163	0.087	0.011	0.008	0.169	0.091	0.012	0.009	0.207	0.111	0.014	0.010	0.316	0.178	0.018	0.015
	MSE	0.155	0.082	0.010	0.008	0.163	0.088	0.012	0.008	0.170	0.091	0.012	0.009	0.208	0.112	0.015	0.010	0.317	0.179	0.020	0.016
	CP	0.947	0.957	0.876	0.953	0.946	0.957	0.873	0.948	0.941	0.947	0.859	0.956	0.928	0.938	0.845	0.944	0.907	0.915	0.810	0.969
EM-MAR(MCAR)	AE	-1.026	1.015	0.179	0.207	-1.025	1.017	0.176	0.207	-1.027	1.017	0.174	0.205	-1.028	1.022	0.172	0.209	-1.028	1.041	0.157	0.212
	RB	0.026	0.015	-0.105	0.035	0.025	0.017	-0.120	0.034	0.027	0.017	-0.128	0.025	0.028	0.022	-0.140	0.045	0.028	0.041	-0.215	0.060
	VAR	0.154	0.082	0.010	0.008	0.151	0.081	0.011	0.006	0.153	0.081	0.011	0.006	0.152	0.080	0.010	0.006	0.151	0.080	0.009	0.006
	MSE	0.155	0.082	0.010	0.008	0.152	0.081	0.012	0.006	0.154	0.081	0.011	0.006	0.153	0.081	0.010	0.006	0.151	0.082	0.011	0.006
	CP	0.947	0.957	0.876	0.953	0.945	0.957	0.874	0.948	0.942	0.947	0.858	0.956	0.928	0.939	0.845	0.942	0.905	0.916	0.810	0.968
EM-MNAR	AE	-1.026	1.015	0.179	0.207	-1.025	1.017	0.176	0.207	-1.027	1.017	0.174	0.205	-1.028	1.022	0.172	0.209	-1.028	1.041	0.157	0.212
	RB	0.026	0.015	-0.105	0.035	0.025	0.017	-0.120	0.034	0.027	0.017	-0.128	0.025	0.028	0.022	-0.140	0.045	0.028	0.041	-0.215	0.060
	VAR	0.154	0.082	0.010	0.008	0.151	0.081	0.011	0.006	0.153	0.081	0.011	0.006	0.152	0.080	0.010	0.006	0.151	0.080	0.009	0.006
	MSE	0.155	0.082	0.010	0.008	0.152	0.081	0.012	0.006	0.154	0.081	0.011	0.006	0.153	0.081	0.010	0.006	0.151	0.082	0.011	0.006
	CP	0.947	0.957	0.876	0.953	0.945	0.957	0.874	0.948	0.942	0.947	0.858	0.956	0.928	0.939	0.845	0.942	0.905	0.916	0.810	0.968

CC, complete case analysis.

Table 3.8: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters for complete data and data under MCAR with one covariate, data simulated from BB(β_0, β_1, ϕ, ω), based on 1000 simulation runs ($n = 50$).*

		Complete data				5% missing				10% missing				25% missing				50% missing			
Method	Quantity	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$
n=50	AE	-1.021	1.018	0.189	0.202	-1.023	1.022	0.189	0.202	-1.024	1.022	0.187	0.204	-1.025	1.022	0.183	0.202	-1.026	1.025	0.176	0.206
	RB	0.021	0.021	-0.055	-0.010	0.023	0.022	-0.055	0.012	0.024	0.022	-0.064	0.019	0.025	0.022	-0.083	0.010	0.026	0.025	-0.120	0.032
CC	VAR	0.091	0.048	0.007	0.005	0.097	0.051	0.007	0.005	0.102	0.054	0.007	0.005	0.124	0.065	0.009	0.006	0.188	0.101	0.013	0.010
	MSE	0.091	0.048	0.007	0.005	0.097	0.051	0.007	0.005	0.103	0.054	0.008	0.005	0.124	0.065	0.009	0.006	0.189	0.102	0.014	0.010
	CP	0.937	0.948	0.888	0.953	0.938	0.947	0.891	0.942	0.926	0.946	0.888	0.945	0.929	0.940	0.876	0.943	0.920	0.931	0.840	0.955
	AE	-1.021	1.018	0.189	0.202	-1.022	1.021	0.190	0.201	-1.024	1.021	0.187	0.203	-1.026	1.022	0.183	0.203	-1.026	1.024	0.177	0.205
	RB	0.021	0.021	-0.055	-0.010	0.022	0.021	-0.050	0.005	0.024	0.021	-0.064	0.015	0.026	0.022	-0.083	0.015	0.026	0.024	-0.115	0.025
EM-MAR(MCAR)	VAR	0.091	0.048	0.007	0.005	0.097	0.051	0.007	0.005	0.102	0.054	0.007	0.005	0.124	0.065	0.009	0.006	0.188	0.101	0.013	0.010
	MSE	0.091	0.048	0.007	0.005	0.097	0.051	0.007	0.005	0.103	0.054	0.008	0.005	0.124	0.065	0.009	0.006	0.189	0.102	0.014	0.010
	CP	0.937	0.948	0.888	0.953	0.939	0.947	0.891	0.941	0.925	0.946	0.888	0.946	0.929	0.941	0.876	0.944	0.923	0.931	0.840	0.954
	AE	-1.021	1.018	0.189	0.202	-1.022	1.021	0.190	0.201	-1.024	1.021	0.187	0.203	-1.026	1.022	0.183	0.203	-1.026	1.024	0.177	0.205
	RB	0.021	0.021	-0.055	-0.010	0.022	0.021	-0.050	0.005	0.024	0.021	-0.064	0.015	0.026	0.022	-0.083	0.015	0.026	0.024	-0.115	0.025
EM-MNAR	VAR	0.091	0.048	0.007	0.005	0.097	0.051	0.007	0.005	0.102	0.054	0.007	0.005	0.124	0.065	0.009	0.006	0.188	0.101	0.013	0.010
	MSE	0.091	0.048	0.007	0.005	0.097	0.051	0.007	0.005	0.103	0.054	0.008	0.005	0.124	0.065	0.009	0.006	0.189	0.102	0.014	0.010
	CP	0.937	0.948	0.888	0.953	0.939	0.947	0.891	0.941	0.925	0.946	0.888	0.946	0.929	0.941	0.876	0.944	0.923	0.931	0.840	0.954

CC, complete case analysis.

Table 3.9: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters for complete data and data under MCAR with one covariate, data simulated from BB(β_0, β_1, ϕ, ω), based on 1000 simulation runs ($n = 100$).*

		Complete data				5% missing				10% missing				25% missing				50% missing			
Method	Quantity	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$
n=100	AE	-1.005	1.012	0.196	0.200	-1.008	1.014	0.196	0.200	-1.009	1.015	0.194	0.200	-1.015	1.018	0.195	0.198	-1.025	1.029	0.190	0.200
	RB	0.005	0.012	-0.020	0.000	0.008	0.014	-0.021	0.001	0.009	0.015	-0.028	0.001	0.015	0.018	-0.026	-0.008	0.025	0.029	-0.050	0.001
CC	VAR	0.046	0.024	0.003	0.002	0.049	0.025	0.004	0.002	0.051	0.027	0.004	0.003	0.062	0.032	0.005	0.003	0.094	0.049	0.007	0.005
	MSE	0.046	0.024	0.003	0.002	0.049	0.025	0.004	0.002	0.051	0.027	0.004	0.003	0.062	0.033	0.005	0.003	0.094	0.050	0.007	0.005
	CP	0.947	0.946	0.930	0.944	0.941	0.946	0.923	0.950	0.948	0.947	0.924	0.946	0.936	0.950	0.907	0.945	0.937	0.940	0.879	0.926
	AE	-1.005	1.012	0.196	0.200	-1.007	1.014	0.196	0.200	-1.009	1.014	0.195	0.200	-1.016	1.017	0.195	0.197	-1.024	1.028	0.191	0.200
	RB	0.005	0.012	-0.020	0.000	0.007	0.014	-0.021	0.001	0.009	0.014	-0.025	0.001	0.016	0.017	-0.026	-0.015	0.024	0.028	-0.045	0.001
EM-MAR(MCAR)	VAR	0.046	0.024	0.003	0.002	0.049	0.025	0.004	0.002	0.051	0.027	0.004	0.003	0.062	0.032	0.005	0.003	0.094	0.049	0.007	0.005
	MSE	0.046	0.024	0.003	0.002	0.049	0.025	0.004	0.002	0.051	0.027	0.004	0.003	0.062	0.033	0.005	0.003	0.094	0.050	0.007	0.005
	CP	0.947	0.946	0.930	0.944	0.942	0.946	0.922	0.950	0.948	0.948	0.924	0.947	0.936	0.951	0.907	0.944	0.936	0.941	0.881	0.925
	AE	-1.005	1.012	0.196	0.200	-1.007	1.014	0.196	0.200	-1.009	1.014	0.195	0.200	-1.016	1.017	0.195	0.197	-1.024	1.028	0.191	0.200
	RB	0.005	0.012	-0.020	0.000	0.007	0.014	-0.021	0.001	0.009	0.014	-0.025	0.001	0.016	0.017	-0.026	-0.015	0.024	0.028	-0.045	0.001
EM-MNAR	VAR	0.046	0.024	0.003	0.002	0.049	0.025	0.004	0.002	0.051	0.027	0.004	0.003	0.062	0.032	0.005	0.003	0.094	0.049	0.007	0.005
	MSE	0.046	0.024	0.003	0.002	0.049	0.025	0.004	0.002	0.051	0.027	0.004	0.003	0.062	0.033	0.005	0.003	0.094	0.050	0.007	0.005
	CP	0.947	0.946	0.930	0.944	0.942	0.946	0.922	0.950	0.948	0.948	0.924	0.947	0.936	0.951	0.907	0.944	0.936	0.941	0.881	0.925

CC, complete case analysis.

Table 3.10: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under MAR with one covariate, data simulated from BB(β_0, β_1, ϕ, ω), based on 1000 simulation runs ($n = 30$).*

		Complete data				5% missing				10% missing				25% missing			
Method	Quantity	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$
n=30	AE	-1.026	1.015	0.179	0.207	-1.035	1.048	0.178	0.205	-1.038	1.053	0.177	0.211	-1.042	1.063	0.171	0.218
CC	RB	0.026	0.015	-0.105	0.035	0.035	0.048	-0.110	0.025	0.038	0.053	-0.115	0.055	0.042	0.063	-0.145	0.090
	VAR	0.154	0.082	0.101	0.008	0.166	0.089	0.012	0.008	0.176	0.096	0.012	0.009	0.217	0.120	0.015	0.011
	MSE	0.155	0.082	0.101	0.008	0.069	0.029	0.002	0.001	0.075	0.032	0.002	0.001	0.103	0.046	0.003	0.001
	CP	0.947	0.957	0.876	0.953	0.920	0.916	0.838	0.950	0.908	0.905	0.841	0.926	0.863	0.862	0.772	0.892
EM-MAR(MCAR)	AE	-1.026	1.015	0.179	0.207	-1.020	1.025	0.179	0.204	-1.020	1.026	0.178	0.208	-1.027	1.029	0.176	0.211
	RB	0.026	0.015	-0.105	0.035	0.020	0.025	-0.105	0.021	0.020	0.026	-0.110	0.040	0.027	0.029	-0.119	0.055
	VAR	0.154	0.082	0.101	0.008	0.154	0.083	0.011	0.008	0.153	0.083	0.011	0.008	0.152	0.082	0.010	0.008
	MSE	0.155	0.082	0.101	0.008	0.061	0.024	0.002	0.001	0.060	0.024	0.002	0.001	0.060	0.024	0.002	0.001
	CP	0.947	0.957	0.876	0.953	0.932	0.932	0.850	0.953	0.924	0.931	0.862	0.949	0.921	0.931	0.836	0.936
EM-MNAR	AE	-1.026	1.015	0.179	0.207	-1.020	1.025	0.179	0.204	-1.020	1.026	0.178	0.208	-1.027	1.029	0.176	0.211
	RB	0.026	0.015	-0.105	0.035	0.020	0.025	-0.105	0.021	0.020	0.026	-0.110	0.040	0.027	0.029	-0.119	0.055
	VAR	0.154	0.082	0.101	0.008	0.154	0.083	0.011	0.008	0.153	0.083	0.011	0.008	0.152	0.082	0.010	0.008
	MSE	0.155	0.082	0.101	0.008	0.061	0.024	0.002	0.001	0.060	0.024	0.002	0.001	0.060	0.024	0.002	0.001
	CP	0.947	0.957	0.876	0.953	0.932	0.932	0.850	0.953	0.924	0.931	0.862	0.949	0.921	0.931	0.836	0.936

CC, complete case analysis.

Table 3.11: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under MAR with one covariate, data simulated from BB(β_0, β_1, ϕ, ω), based on 1000 simulation runs ($n = 50$).*

		Complete data				5% missing				10% missing				25% missing			
Method	Quantity	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$
n=50	AE	-1.021	1.018	0.189	0.202	-1.024	1.022	0.187	0.201	-1.026	1.024	0.184	0.202	-1.027	1.027	0.177	0.202
	RB	0.021	0.021	-0.055	-0.010	0.024	0.022	-0.065	0.005	0.026	0.024	-0.080	0.010	0.027	0.027	-0.115	0.010
CC	VAR	0.091	0.048	0.007	0.005	0.097	0.051	0.007	0.005	0.103	0.055	0.008	0.005	0.124	0.068	0.009	0.006
	MSE	0.091	0.048	0.007	0.005	0.031	0.012	0.001	0.000	0.034	0.013	0.001	0.000	0.045	0.018	0.001	0.001
	CP	0.937	0.948	0.888	0.953	0.937	0.926	0.874	0.904	0.924	0.936	0.853	0.903	0.888	0.888	0.812	0.859
	AE	-1.021	1.018	0.189	0.202	-1.020	1.019	0.190	0.200	-1.021	1.021	0.187	0.199	-1.022	1.023	0.183	0.200
	RB	0.021	0.021	-0.055	-0.010	0.020	0.019	-0.051	0.002	0.021	0.021	-0.064	-0.006	0.022	0.023	-0.084	0.002
EM-MAR(MCAR)	VAR	0.091	0.048	0.007	0.005	0.091	0.048	0.007	0.005	0.091	0.048	0.007	0.005	0.089	0.047	0.006	0.004
	MSE	0.091	0.048	0.007	0.005	0.028	0.011	0.001	0.000	0.028	0.011	0.001	0.000	0.027	0.011	0.001	0.000
	CP	0.937	0.948	0.888	0.953	0.940	0.944	0.877	0.917	0.940	0.951	0.865	0.922	0.936	0.943	0.864	0.920
	AE	-1.021	1.018	0.189	0.202	-1.020	1.019	0.190	0.200	-1.021	1.021	0.187	0.199	-1.022	1.023	0.183	0.200
	RB	0.021	0.021	-0.055	-0.010	0.020	0.019	-0.051	0.002	0.021	0.021	-0.064	-0.006	0.022	0.023	-0.084	0.002
EM-MNAR	VAR	0.091	0.048	0.007	0.005	0.091	0.048	0.007	0.005	0.091	0.048	0.007	0.005	0.089	0.047	0.006	0.004
	MSE	0.091	0.048	0.007	0.005	0.028	0.011	0.001	0.000	0.028	0.011	0.001	0.000	0.027	0.011	0.001	0.000
	CP	0.937	0.948	0.888	0.953	0.940	0.944	0.877	0.917	0.940	0.951	0.865	0.922	0.936	0.943	0.864	0.920

CC, complete case analysis.

Table 3.12: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under MAR with one covariate, data simulated from BB(β_0, β_1, ϕ, ω), based on 1000 simulation runs ($n = 100$).*

		Complete data				5% missing				10% missing				25% missing			
Method	Quantity	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$
n=100	AE	-1.005	1.012	0.196	0.200	-1.020	1.014	0.196	0.198	-1.022	1.016	0.195	0.197	-1.029	1.023	0.195	0.197
	RB	0.005	0.012	-0.020	0.000	0.020	0.014	-0.018	-0.010	0.022	0.016	-0.025	-0.017	0.029	0.023	-0.025	-0.015
CC	VAR	0.046	0.024	0.003	0.002	0.048	0.025	0.004	0.002	0.051	0.027	0.004	0.003	0.061	0.032	0.005	0.003
	MSE	0.046	0.024	0.003	0.002	0.011	0.004	0.000	0.000	0.012	0.005	0.000	0.000	0.016	0.006	0.000	0.000
	CP	0.947	0.946	0.930	0.944	0.938	0.942	0.906	0.928	0.923	0.942	0.884	0.919	0.901	0.895	0.865	0.881
	AE	-1.005	1.012	0.196	0.200	-1.008	1.011	0.199	0.200	-1.012	1.013	0.198	0.200	-1.015	1.012	0.197	0.200
	RB	0.005	0.012	-0.020	0.000	0.008	0.011	-0.007	0.001	0.012	0.013	-0.012	-0.002	0.015	0.012	-0.015	-0.001
EM-MAR(MCAR)	VAR	0.046	0.024	0.003	0.002	0.046	0.024	0.003	0.002	0.045	0.024	0.003	0.002	0.045	0.023	0.003	0.002
	MSE	0.046	0.024	0.003	0.002	0.010	0.004	0.000	0.000	0.010	0.004	0.000	0.000	0.010	0.004	0.000	0.000
	CP	0.947	0.946	0.930	0.944	0.949	0.950	0.910	0.933	0.934	0.949	0.901	0.935	0.949	0.942	0.899	0.935
	AE	-1.005	1.012	0.196	0.200	-1.008	1.011	0.199	0.200	-1.012	1.013	0.198	0.200	-1.015	1.012	0.197	0.200
	RB	0.005	0.012	-0.020	0.000	0.008	0.011	-0.007	0.001	0.012	0.013	-0.012	-0.002	0.015	0.012	-0.015	-0.001
EM-MNAR	VAR	0.046	0.024	0.003	0.002	0.046	0.024	0.003	0.002	0.045	0.024	0.003	0.002	0.045	0.023	0.003	0.002
	MSE	0.046	0.024	0.003	0.002	0.010	0.004	0.000	0.000	0.010	0.004	0.000	0.000	0.010	0.004	0.000	0.000
	CP	0.947	0.946	0.930	0.944	0.949	0.950	0.910	0.933	0.934	0.949	0.901	0.935	0.949	0.942	0.899	0.935

CC, complete case analysis.

Table 3.13: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under MNAR with one covariate, data simulated from BB(β_0, β_1, ϕ, ω), based on 1000 simulation runs ($n = 30$).*

		Complete data				5% missing				10% missing				25% missing			
Method	Quantity	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$
n=30	AE	-1.026	1.015	0.179	0.207	-1.057	1.037	0.182	0.208	-1.065	1.031	0.176	0.211	-1.110	1.037	0.174	0.228
	RB	0.026	0.015	-0.105	0.035	0.057	0.037	-0.090	0.040	0.065	0.031	-0.120	0.055	0.110	0.037	-0.130	0.140
CC	VAR	0.154	0.082	0.101	0.008	0.171	0.091	0.012	0.008	0.183	0.099	0.013	0.009	0.235	0.132	0.015	0.013
	MSE	0.155	0.082	0.101	0.008	0.174	0.092	0.013	0.009	0.187	0.100	0.013	0.010	0.247	0.134	0.016	0.013
	CP	0.947	0.957	0.876	0.953	0.900	0.904	0.848	0.946	0.873	0.898	0.822	0.942	0.836	0.834	0.765	0.910
	AE	-1.026	1.015	0.179	0.207	-1.056	1.037	0.183	0.208	-1.067	-1.030	0.176	0.211	-1.110	1.037	0.173	0.228
	RB	0.026	0.015	-0.105	0.035	0.056	0.037	-0.085	0.040	0.067	-2.030	-0.120	0.055	0.110	0.037	-0.135	0.140
EM-MAR(MCAR)	VAR	0.154	0.082	0.101	0.008	0.156	0.082	0.011	0.008	0.153	0.080	0.011	0.008	0.150	0.077	0.010	0.008
	MSE	0.155	0.082	0.101	0.008	0.159	0.083	0.012	0.008	0.157	4.201	0.011	0.008	0.162	0.079	0.010	0.009
	CP	0.947	0.957	0.876	0.953	0.898	0.903	0.848	0.946	0.872	0.898	0.811	0.943	0.837	0.834	0.765	0.908
	AE	-1.026	1.015	0.179	0.207	-1.027	1.015	0.193	0.203	-1.028	1.016	0.193	0.203	-1.030	1.018	0.191	0.205
	RB	0.026	0.015	-0.105	0.035	0.027	0.015	-0.035	0.015	0.028	0.016	-0.035	0.015	0.030	0.018	-0.045	0.025
EM-MNAR	VAR	0.154	0.082	0.101	0.008	0.154	0.082	0.011	0.008	0.153	0.081	0.011	0.007	0.147	0.080	0.010	0.007
	MSE	0.155	0.082	0.101	0.008	0.155	0.082	0.011	0.008	0.154	0.081	0.011	0.007	0.148	0.080	0.010	0.007
	CP	0.947	0.957	0.876	0.953	0.918	0.919	0.858	0.955	0.906	0.922	0.855	0.960	0.904	0.927	0.822	0.949

CC, complete case analysis.

Table 3.14: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under MNAR with one covariate, data simulated from BB(β_0, β_1, ϕ, ω), based on 1000 simulation runs ($n = 50$).*

		Complete data				5% missing				10% missing				25% missing			
Method	Quantity	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$
n=50	AE	-1.021	1.018	0.189	0.202	-1.023	1.018	0.189	0.207	-1.024	1.021	0.189	0.207	-1.026	1.025	0.185	0.220
	RB	0.021	0.021	-0.055	-0.010	0.023	0.018	-0.055	0.035	0.024	0.021	-0.055	0.035	0.026	0.025	-0.075	0.100
CC	VAR	0.091	0.048	0.007	0.005	0.102	0.054	0.007	0.005	0.110	0.059	0.008	0.006	0.143	0.080	0.011	0.008
	MSE	0.091	0.048	0.007	0.005	0.102	0.054	0.008	0.005	0.111	0.059	0.008	0.006	0.144	0.080	0.011	0.008
	CP	0.937	0.948	0.888	0.953	0.932	0.926	0.875	0.936	0.922	0.912	0.873	0.925	0.863	0.848	0.808	0.888
	AE	-1.021	1.018	0.189	0.202	-1.022	1.018	0.190	0.207	-1.024	1.021	0.190	0.207	-1.026	1.025	0.186	0.220
	RB	0.021	0.021	-0.055	-0.010	0.022	0.018	-0.050	0.035	0.024	0.021	-0.050	0.035	0.026	0.025	-0.070	0.100
EM-MAR(MCAR)	VAR	0.091	0.048	0.007	0.005	0.091	0.047	0.007	0.004	0.087	0.045	0.007	0.005	0.078	0.038	0.006	0.004
	MSE	0.091	0.048	0.007	0.005	0.091	0.047	0.007	0.005	0.088	0.045	0.007	0.005	0.079	0.039	0.006	0.005
	CP	0.937	0.948	0.888	0.953	0.931	0.925	0.876	0.936	0.921	0.913	0.874	0.925	0.865	0.847	0.808	0.882
	AE	-1.021	1.018	0.189	0.202	-1.021	1.017	0.196	0.201	-1.021	1.019	0.194	0.201	-1.024	1.015	0.190	0.203
	RB	0.021	0.021	-0.055	-0.010	0.021	0.017	-0.020	0.005	0.021	0.019	-0.030	0.005	0.024	0.015	-0.050	0.015
EM-MNAR	VAR	0.091	0.048	0.007	0.005	0.093	0.049	0.007	0.005	0.092	0.049	0.007	0.005	0.090	0.048	0.007	0.004
	MSE	0.091	0.048	0.007	0.005	0.093	0.049	0.007	0.005	0.093	0.049	0.007	0.005	0.091	0.048	0.007	0.004
	CP	0.937	0.948	0.888	0.953	0.941	0.940	0.882	0.950	0.943	0.938	0.884	0.950	0.944	0.948	0.866	0.936

CC, complete case analysis.

Table 3.15: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under MNAR with one covariate, data simulated from BB(β_0, β_1, ϕ, ω), based on 1000 simulation runs ($n = 100$).*

		Complete data				5% missing				10% missing				25% missing			
Method	Quantity	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\phi = 0.2$	$\omega = 0.2$
n=100	AE	-1.005	1.012	0.196	0.200	-1.013	1.011	0.193	0.205	-1.017	1.013	0.192	0.206	-1.023	1.017	0.190	0.215
	RB	0.005	0.012	-0.020	0.000	0.013	0.011	-0.035	0.025	0.017	0.013	-0.040	0.030	0.023	0.017	-0.050	0.075
CC	VAR	0.046	0.024	0.003	0.002	0.050	0.026	0.004	0.003	0.054	0.028	0.004	0.003	0.070	0.037	0.005	0.004
	MSE	0.046	0.024	0.003	0.002	0.050	0.026	0.004	0.003	0.055	0.028	0.004	0.003	0.071	0.038	0.005	0.004
	CP	0.947	0.946	0.930	0.944	0.923	0.934	0.906	0.931	0.908	0.907	0.903	0.919	0.851	0.841	0.833	0.885
	AE	-1.005	1.012	0.196	0.200	-1.012	1.011	0.193	0.204	-1.017	1.011	0.193	0.206	-1.022	1.016	0.191	0.215
	RB	0.005	0.012	-0.020	0.000	0.012	0.011	-0.035	0.020	0.017	0.011	-0.035	0.030	0.022	0.016	-0.045	0.075
EM-MAR(MCAR)	VAR	0.046	0.024	0.003	0.002	0.042	0.021	0.003	0.002	0.038	0.019	0.003	0.002	0.028	0.011	0.002	0.002
	MSE	0.046	0.024	0.003	0.002	0.042	0.021	0.003	0.002	0.039	0.019	0.003	0.002	0.028	0.012	0.002	0.002
	CP	0.947	0.946	0.930	0.944	0.924	0.930	0.907	0.930	0.907	0.906	0.904	0.919	0.852	0.840	0.833	0.883
	AE	-1.005	1.012	0.196	0.200	-1.007	1.003	0.195	0.200	-1.008	1.002	0.194	0.201	-1.014	1.002	0.191	0.202
	RB	0.005	0.012	-0.020	0.000	0.007	0.003	-0.025	0.000	0.008	0.002	-0.030	0.005	0.014	0.002	-0.045	0.010
EM-MNAR	VAR	0.046	0.024	0.003	0.002	0.046	0.024	0.003	0.002	0.046	0.023	0.003	0.002	0.045	0.023	0.003	0.002
	MSE	0.046	0.024	0.003	0.002	0.046	0.024	0.003	0.002	0.046	0.023	0.003	0.002	0.046	0.023	0.003	0.002
	CP	0.947	0.946	0.930	0.944	0.943	0.944	0.913	0.937	0.938	0.936	0.923	0.927	0.929	0.922	0.890	0.921

CC, complete case analysis.

Table 3.16: *The number of females with* $0, 1, 2,$ *etc. dead implants when 5-10 zygotes were implanted after matings during the first* 7 *days after irradiation of males with* 0 *(control), 300 R and 600 R.*

Dose		Dead Implants									
(R)	Implantations	0	1	2	3	4	5	6	7	Total	dead
0	5	30	27	9	5	_	_	_	_	71	16.9
	6	86	51	14	4	1	_	_	_	156	10.1
	7	111	73	31	8	1	_	_	_	224	10.4
	8	79	44	23	3	_	1	_	_	150	8.7
	9	32	29	8	1	_	_	_	_	70	7.6
	10	5	5	2	_	_	_	_	_	12	7.5
300	5	27	41	32	17	4	_	_	_	121	28.4
	6	28	47	59	28	6	1	1	_	170	27.8
	7	31	61	54	20	19	1	_	_	186	23.8
	8	12	32	24	22	8	1	_	_	99	23.1
	9	1	6	9	6	1	1	_	_	24	23.6
	10	1	2	1	_	_	_	_	_	4	10.0
600	5	16	32	48	49	15	_	_	_	160	41.9
	6	7	35	45	37	20	9	_	_	153	39.3
	7	5	22	27	36	17	9	3	1	120	38.3
	8	1	4	12	11	8	7	_	2	45	39.4
	9	_	_	2	2	2	_	1	_	7	38.1
	10	_	_	_	_	_	_	_	1	1	70.0

Table 3.17: *Estimates and standard error of the parameters for mutagenic data under the three missing data mechanism.*

(a) MCAR

Method	Quantity	Complete data			5% missing			10% missing			25% missing		
		β_0	β_1	ϕ	β_0	β_1	ϕ	β_0	β_1	ϕ	β_0	β_1	ϕ
CC	MLE	-1.313	0.702	0.026	-1.329	0.714	0.023	-1.303	0.700	0.027	-1.302	0.687	0.023
	VAR	0.025	0.025	0.007	0.026	0.026	0.007	0.028	0.026	0.007	0.029	0.029	0.008
EM-MAR(MCAR)	MLE	-1.313	0.702	0.026	-1.327	0.714	0.024	-1.305	0.701	0.027	-1.304	0.691	0.023
	VAR	0.025	0.025	0.007	0.024	0.024	0.006	0.027	0.024	0.006	0.028	0.029	0.007
EM-MNAR	MLE	-1.313	0.702	0.026	-1.327	0.714	0.024	-1.305	0.701	0.027	-1.304	0.691	0.023
	VAR	0.025	0.025	0.007	0.024	0.024	0.006	0.027	0.024	0.006	0.028	0.029	0.007

(b) MAR

Method	Quantity	Complete data			5% missing			10% missing			25% missing		
		β_0	β_1	ϕ	β_0	β_1	ϕ	β_0	β_1	ϕ	β_0	β_1	ϕ
CC	MLE	-1.313	0.702	0.026	-1.300	0.707	0.027	-1.299	0.718	0.030	-1.356	0.635	0.027
	VAR	0.025	0.025	0.007	0.027	0.026	0.007	0.028	0.027	0.007	0.030	0.028	0.008
EM-MAR(MCAR)	MLE	-1.313	0.702	0.026	-1.313	0.704	0.027	-1.317	0.711	0.026	-1.322	0.713	0.025
	VAR	0.025	0.025	0.007	0.025	0.025	0.007	0.026	0.025	0.007	0.027	0.026	0.007
EM-MNAR	MLE	-1.313	0.702	0.026	-1.313	0.704	0.027	-1.317	0.711	0.026	-1.322	0.713	0.025
	VAR	0.025	0.025	0.007	0.025	0.025	0.007	0.026	0.025	0.007	0.027	0.026	0.007

(c) MNAR

Method	Quantity	Complete data			5% missing			10% missing			25% missing		
		β_0	β_1	ϕ	β_0	β_1	ϕ	β_0	β_1	ϕ	β_0	β_1	ϕ
CC	MLE	-1.313	0.702	0.026	-1.301	0.681	0.041	-1.326	0.644	0.047	-1.500	0.531	0.027
	VAR	0.025	0.025	0.007	0.028	0.027	0.008	0.029	0.028	0.008	0.030	0.029	0.008
EM-MAR(MCAR)	MLE	-1.313	0.702	0.026	-1.302	0.682	0.042	-1.325	0.646	0.045	-1.402	0.537	0.027
	VAR	0.025	0.025	0.007	0.027	0.027	0.008	0.028	0.028	0.008	0.029	0.028	0.008
EM-MNAR	MLE	-1.313	0.702	0.026	-1.301	0.685	0.032	-1.318	0.659	0.039	-1.353	0.592	0.024
	VAR	0.025	0.025	0.007	0.027	0.026	0.007	0.027	0.027	0.008	0.027	0.025	0.007

CC, complete case analysis.

Table 3.18: *The number of females with* $0, 1, 2,$ *etc. dead implants when 5-10 zygotes were implanted after matings during the first* 7 *days after irradiation of males with* 0 *(control), 300 R and 600 R (Generated from Table 3.16).*

Dose		Dead Implants									
(R)	Implantations	0	1	2	3	4	5	6	7	Total	dead
0	5	47	17	7	0	_	_	_	_	71	8.7
	6	86	45	22	2	1	1	_	_	156	10.6
	7	94	76	36	14	4	_	_	_	224	13.1
	8	64	49	27	10	_	_	_	_	150	11.1
	9	31	24	11	1	3	_	_	_	70	9.7
	10	4	5	3	_	_	_	_	_	12	9.2
300	5	42	40	28	10	1	_	_	_	121	21.5
	6	48	60	42	13	4	3	_	_	170	20.1
	7	37	56	50	36	5	2	_	_	186	22.6
	8	20	25	27	17	8	1	1	_	99	21.8
	9	2	6	9	3	_	3	1	_	24	25.0
	10	1	_	2	0	1	_	_	_	4	20.0
600	5	13	36	62	43	4	2	_	_	160	39.4
	6	13	20	53	43	16	7	1	_	153	39.2
	7	6	15	32	26	21	16	4	_	120	41.1
	8	2	6	7	20	7	2	1	_	45	34.4
	9	1	0	1	1	2	0	2	_	7	39.7
	10	1	_	_	_	_	_	_	_	1	0

Table 3.19: *Estimates and standard error of the parameters for new mutagenic data under the three missing data mechanism.*

(a) MCAR

		Complete data				5% missing				10% missing				25% missing			
Method	Quantity	β_0	β_1	ϕ	ω	β_0	β_1	ϕ	ω	β_0	β_1	ϕ	ω	β_0	β_1	ϕ	ω
CC	MLE	-1.307	0.660	0.020	0.033	-1.293	0.661	0.011	0.045	-1.325	0.678	0.024	0.024	-1.316	0.659	0.016	0.029
	VAR	0.028	0.024	0.001	0.011	0.031	0.025	0.008	0.012	0.031	0.026	0.008	0.013	0.034	0.029	0.009	0.013
EM-MAR(MCAR)	MLE	-1.307	0.660	0.020	0.032	-1.294	0.661	0.012	0.044	-1.323	0.678	0.026	0.025	-1.316	0.660	0.016	0.028
	VAR	0.028	0.024	0.001	0.011	0.030	0.024	0.008	0.011	0.031	0.025	0.008	0.012	0.033	0.028	0.009	0.013
EM-MNAR	MLE	-1.307	0.660	0.020	0.032	-1.294	0.661	0.012	0.044	-1.323	0.678	0.026	0.025	-1.316	0.660	0.016	0.028
	VAR	0.028	0.024	0.001	0.011	0.030	0.024	0.008	0.011	0.031	0.025	0.008	0.012	0.033	0.028	0.009	0.013

(b) MAR

		Complete data				5% missing				10% missing				25% missing			
Method	Quantity	β_0	β_1	ϕ	ω	β_0	β_1	ϕ	ω	β_0	β_1	ϕ	ω	β_0	β_1	ϕ	ω
CC	MLE	-1.307	0.660	0.020	0.033	-1.303	0.654	0.018	0.036	-1.205	0.783	0.024	0.060	-1.083	0.660	0.000	0.103
	VAR	0.028	0.024	0.001	0.011	0.030	0.026	0.008	0.012	0.031	0.027	0.008	0.013	0.033	0.029	0.008	0.014
EM-MAR(MCAR)	MLE	-1.307	0.660	0.020	0.033	-1.304	0.655	0.019	0.034	-1.302	0.650	0.021	0.029	-1.317	0.663	0.022	0.020
	VAR	0.028	0.024	0.001	0.011	0.028	0.025	0.080	0.012	0.030	0.025	0.008	0.012	0.030	0.024	0.008	0.013
EM-MNAR	MLE	-1.307	0.660	0.020	0.033	-1.304	0.655	0.019	0.034	-1.302	0.650	0.021	0.029	-1.317	0.663	0.022	0.020
	VAR	0.028	0.024	0.001	0.011	0.028	0.025	0.080	0.012	0.030	0.025	0.008	0.012	0.030	0.024	0.008	0.013

(c) MNAR

		Complete data				5% missing				10% missing				25% missing			
Method	Quantity	β_0	β_1	ϕ	ω	β_0	β_1	ϕ	ω	β_0	β_1	ϕ	ω	β_0	β_1	ϕ	ω
CC	MLE	-1.307	0.660	0.020	0.033	-1.315	0.661	0.022	0.029	-1.268	0.726	0.018	0.042	-1.317	0.651	0.019	0.038
	VAR	0.028	0.024	0.001	0.011	0.031	0.025	0.008	0.012	0.029	0.027	0.008	0.012	0.029	0.027	0.008	0.012
EM-MAR(MCAR)	MLE	-1.307	0.660	0.020	0.033	-1.314	0.662	0.023	0.029	-1.268	0.727	0.017	0.040	-1.315	0.647	0.020	0.038
	VAR	0.028	0.024	0.001	0.011	0.030	0.025	0.008	0.012	0.029	0.027	0.008	0.012	0.029	0.027	0.008	0.012
EM-MNAR	MLE	-1.307	0.660	0.020	0.033	-1.311	0.662	0.022	0.029	-1.276	0.722	0.020	0.038	-1.312	0.657	0.020	0.035
	VAR	0.028	0.024	0.001	0.011	0.027	0.026	0.007	0.012	0.029	0.025	0.007	0.011	0.029	0.025	0.007	0.011

CC, complete case analysis.

Chapter 4

Estimation for Zero-Inflated Beta-Binomial Regression Model with Covariate Measurement Error And/or Missing Responses

4.1 Introduction

In chapter 3, we developed estimation procedures for the parameters of a zero-inflated beta-binomial regression model with missing data. The purpose of this chapter is to develop inference procedures for the parameters of a zero-inflated beta-binomial model where information on some of the covariates is recorded with errors and/or some observations of the binomial responses may be missing. A weighted expectation maximization algorithm (Dempster et al. (1977)) is developed for the maximum likelihood (ML) estimation of the parameters involved. Extensive simulations are

conducted to study the properties of the estimates using different measures, such as, average estimates (AE), relative bias (RB), variance(VAR), mean squared error (MSE) and coverage probability (CP) of estimates. Simulations show much superior properties of the estimates obtained using the weighted expectation maximization algorithm. Some illustrative examples and a discussion are given.

The zero-inflated beta-binomial model is introduced in Section 2. In this section we also develop a procedure for the estimation of the parameters. Results of an extensive simulation study are reported in Section 3. Some illustrative examples are given in Section 4 and a discussion leading to some conclusions is given in Section 5.

4.2 The zero-inflated beta-binomial model and estimation procedure

4.2.1 The zero-inflated beta-binomial model

For a quick introduction to the proposed method it is appropriate to present the standard form for the ZIBB model. For a particular litter i, given m_i, the number of live foetuses in the litter, y_i, the number of foetuses affected, is a random variable having a beta-binomial distribution with parameters α and β, i.e,

$$f(y_i;\alpha,\beta)=\binom{m_i}{y_i}B(\alpha+y_i,m_i+\beta-y_i)/B(\alpha,\beta). \tag{4.1}$$

If $\pi=\frac{\alpha}{\alpha+\beta}$, and $\phi=\frac{1}{\alpha+\beta}$, we have

$$f(y_i;\alpha,\beta)=\binom{m_i}{y_i}\frac{\prod_{r=0}^{y_i-1}(\pi+r\phi)\prod_{r=0}^{m_i-y_i-1}(1-\pi+r\phi)}{\prod_{r=0}^{m_i-1}(1+r\phi)}, \tag{4.2}$$

with $E(Y_i) = m_i\pi$ and $Var(Y_i) = m_i\pi(1-\pi)\big[1 + \frac{(m_i-1)\phi}{1+\phi}\big]$. We denote the beta-binomial distribution as $BB(\pi, \phi)$. As $\phi \to 0$, $BB(\pi, \phi)$ tends to the binomial (π) distribution and for $\phi = 0$ we have $Var(Y_i) = m_i\pi(1-\pi)$ and the $BB(\pi, \phi)$ becomes the binomial (π) distribution.

The zero-inflated beta binomial regression model (Deng and Paul (2005)) can be written as

$$f(y_i|x_i;\pi,\phi,\omega) = \begin{cases} \omega + (1-\omega)\dfrac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi)}{\prod_{r=0}^{m_i-1}(1+r\phi)} & \text{if } y_i = 0, \\ (1-\omega)\binom{m_i}{y_i}\dfrac{\prod_{r=0}^{y_i-1}(\pi+r\phi)\prod_{r=0}^{m_i-y_i-1}(1-\pi+r\phi)}{\prod_{r=0}^{m_i-1}(1+r\phi)} & \text{if } y_i > 0, \end{cases} \tag{4.3}$$

with $E(Y_i) = (1-\omega)m_i\pi$, and $Var(Y_i) = (1-\omega)m_i\pi(1-\pi)\frac{1+m_i\phi}{1+\phi} + (1-\omega)\omega m_i^2\pi^2$, where ω is the zero-inflation parameter. We denote this distribution by $ZIBB(\pi, \phi, \omega)$ distribution.

4.2.2 The estimation procedure

Suppose data from the $ZIBB(\pi, \phi, \omega)$ model for the i^{th} litter are (y_i, x_i), given the number m_i of litter size, $i = 1, \ldots, n$, y_i represents the response variable and x_i represents a $p \times 1$ vector of covariates with the regression parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, such that $\pi_i = \exp(\sum_{j=0}^{p} x_{ij}\beta_j)/(1 + \exp(\sum_{j=0}^{p} x_{ij}\beta_j))$. Here β_0 is the intercept parameter in which case $x_{i0} = 1$ for all i.

4.2.2.1 Estimation of ψ with no covariate measurement error

For data without covariate measurement error, the log likelihood, apart from a constant, using the probability mass function given in equation (4.3), can be written as

$$\begin{aligned} l(\beta_j, \phi, \gamma | y_i) = \sum_{i=1}^{n} \Bigg[& -\log(1+\gamma) + \log\left[\gamma + \frac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1}(1+r\phi)}\right] I_{\{y_i=0\}} \\ & + \Big[\sum_{r=0}^{y_i-1}\log(\pi_i+r\phi) + \sum_{r=0}^{m_i-y_i-1}\log(1-\pi_i+r\phi) \\ & - \sum_{r=0}^{m_i-1}\log(1+r\phi)\Big] I_{\{y_i>0\}}\Bigg], \end{aligned} \tag{4.4}$$

where $\gamma = \omega/(1-\omega)$. Note, γ transforms the space of ω from $(0,1)$ onto $(0,\infty)$ which makes optimization of l easier (Deng and Paul (2005)). Let $\psi = (\beta, \phi, \gamma)$. Then the maximum likelihood estimates of the parameters ψ can be obtained by simultaneously solving the following estimating equations

$$\begin{aligned} \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \Bigg[& \left[\frac{\Big(-\sum_{j=0}^{m_i-1}\prod_{r=0, r\neq j}^{m_i-1}(1+r\phi-\pi_i)\Big)}{\prod_{r=0}^{m_i-1}(1+r\phi)\left(\gamma + \frac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1}(1+r\phi)}\right)}\right] I_{\{y_i=0\}} \\ & + \left[\sum_{r=0}^{y_i-1}\frac{1}{\pi_i+r\phi} - \sum_{r=0}^{m_i-y_i-1}\frac{1}{1-\pi_i+r\phi}\right] I_{\{y_i>0\}}\Bigg] \frac{\partial \pi_i}{\partial \beta_j} = 0, \end{aligned}$$

$$\frac{\partial l}{\partial \phi}=\sum_{i=1}^{n}\Bigg[\Bigg[\frac{\Big(\sum_{j=0}^{m_i-1} j\prod_{r=0,r\neq j}^{m_i-1}(1+r\phi-\pi_i)\Big)\prod_{r=0}^{m_i-1}(1+r\phi)}{\Big(\prod_{r=0}^{m_i-1}(1+r\phi)\Big)^2\Big(\gamma+\dfrac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1}(1+r\phi)}\Big)}$$
$$-\frac{\Big(\sum_{j=0}^{m_i-1} j\prod_{r=0,r\neq j}^{m_i-1}(1+r\phi)\Big)\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)}{\Big(\prod_{r=0}^{m_i-1}(1+r\phi)\Big)^2\Big(\gamma+\dfrac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1}(1+r\phi)}\Big)}\Bigg]I_{\{y_i=0\}}$$
$$+\Bigg[\sum_{r=0}^{y_i-1}\frac{r}{\pi_i+r\phi}+\sum_{r=0}^{m_i-y_i-1}\frac{r}{1-\pi_i+r\phi}-\sum_{r=0}^{m_i-1}\frac{r}{1+r\phi}\Bigg]I_{\{y_i>0\}}\Bigg]=0$$

and

$$\frac{\partial l}{\partial \gamma}=\sum_{i=1}^{n}\Bigg[-(1+\gamma)^{-1}+\Bigg(\gamma+\frac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1}(1+r\phi)}\Bigg)^{-1}I_{\{y_i=0\}}\Bigg]=0,$$

where $\dfrac{\partial \pi_i}{\partial \beta_j}=X_{ij}\exp(\sum_{j=0}^{p}X_{ij}\beta_j)/(1+exp(\sum_{j=0}^{p}X_{ij}\beta_j))^2$. Denote these estimates by $\hat{\psi}$. The observed information matrix of $\hat{\psi}$ is given by

$$H_0=-Q''(\psi,\alpha|\psi^{(s)})=\ -\sum_{i=1}^{n}\frac{\partial^2}{\partial\psi\partial\psi'}l_i(\psi;y_i,x_i|\hat{\psi}). \tag{4.5}$$

The elements of this matrix are given in appendix 2. Of course, if it is convenient, these parameters can also be estimated by directly maximizing the log-likelihood function (4.4). However, in practice, through tests (Deng and Paul (2005)), if it is found that the zero-inflation parameter is insignificant, then data analysis should be based on the beta-binomial model (4.2). The parameters β_j and ϕ can be estimated by

solving the estimating equations given in Appendix 1. The elements of the observed information matrix corresponding to likelihood (4.2) are also given in this appendix.

4.2.2.2 Estimation of the parameters with covariate measurement error

We partition the vector of covariates x_i for the ith observation as (u_i, z_i), the vector u_i is observed only indirectly through the measurement w_i and z_i is observed without error. Note that u_i and w_i are q dimensional while z_i is $p-q$ dimensional.

Following Carroll et al. (2006), the measurement error model can be classified into two general types which are used to relate w_i to u_i:

1. Error model, which includes classical measurement error model.

$$w_i = \tau_0 + \tau_u u_i + \tau_z z_i + e_i. \tag{4.6}$$

The error term e_i is independent of u_i, z_i and the responses and it is often assumed that e_i has mean zero and it follows a known distribution $f(0, \Sigma)$, where Σ is the covariance matrix of e_i. The intercept τ_0 is a vector, which can be written as $\tau_0 = (\tau_{01}, \cdots \tau_{0q})^T$. The coefficients $\tau_u = (\tau_{u1}, \cdots \tau_{uq})^T$, $\tau_z = (\tau_{z1}, \cdots \tau_{zq})^T$, where $\tau_{uj} (j = 1, \cdots, q)$ and $\tau_{zk} (k = 1, \cdots, p-q)$ are $q \times 1$ and $(p-q) \times 1$ vectors respectively. If we set $\tau_0 = \mathbf{0}$, $\tau_z = \mathbf{0}$, and $\tau_u = I_q$, where I_q is the $q \times q$ dimensional identity matrix, we have the classical measurement error model.

2. Regression calibration model, which includes the Berkson error model.

$$u_i = \tau_0 + \tau_w w_i + \tau_z z_i + e_i, \tag{4.7}$$

where, $\tau_w = (\tau_{w1}, \cdots \tau_{wq})^T$. If we set $\tau_0 = \mathbf{0}$, $\tau_z = \mathbf{0}$, and $\tau_w = I_q$, we have the Berkson measurement error model.

Which of these models should be applied to the analysis of binomial data? See for example, the data in Table 4.13 in which all individuals in a small group are given the same dose. However, because of the size of the implants the actual dose will vary from animal to animal. In this situation the Berkson model is appropriate (see Carroll et al. (2006, p 27)).

We suppose $f(y_i|w_i, u_i, z_i) = f(y_i|u_i, z_i)$, which is called nondifferential error mechanism (Carroll et al. (2006)). That is, w are statistically independent of y given (u, z). Now, the true values of the u are not available. So, we treat them as missing data and obtain maximum likelihood estimates of the parameters involved by using the EM algorithm (Schafer (1987)).

The likelihood for the complete data with error prone covariates u_i and error free covariate z_i is

$$\prod_{i=1}^{n} f(y_i|u_i, z_i; \psi) f(u_i|w_i, z_i; \tau, \Sigma). \tag{4.8}$$

The log-likelihood for subject i is

$$\begin{aligned} l_i = \;& logf(y_i|u_i, z_i; \psi) + logf(u_i|w_i, z_i; \tau, \Sigma) \\ = \;& -\log(1+\gamma) + \log\left[\gamma + \frac{\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)}{\prod_{r=0}^{m_i-1}(1+r\phi)}\right] I_{\{y_i=0\}} \\ & + \Big[\sum_{r=0}^{y_i-1}\log(\pi_i+r\phi) + \sum_{r=0}^{m_i-y_i-1}\log(1-\pi_i+r\phi) - \sum_{r=0}^{m_i-1}\log(1+r\phi)\Big] I_{\{y_i>0\}} \\ & + logf(u_i|w_i, z_i; \tau, \Sigma). \end{aligned} \tag{4.9}$$

Denote $\Gamma = (\psi, \tau, \Sigma)$. The E-step requires the calculation of the conditional expectation of (4.9) with respect to u_i given the observed data and current estimates

of the parameters. We have observed data (y_i, w_i, z_i). The E-Step is given as

$$\begin{aligned} Q_i(\psi, \tau, \Sigma|\Gamma^{(t)}) = & \ E\left[l_i|y_i, w_i, z_i; \Gamma^{(t)}\right] \\ = & \int [logf(y_i|u_i, z_i; \psi) + logf(u_i|w_i, z_i; \tau, \Sigma)] \\ & \cdot f(u_i|y_i, w_i, z_i; \Gamma^{(t)})du_i. \end{aligned} \tag{4.10}$$

Since the above integration has no closed form, we use the Monte Carlo (MC) version of the EM algorithm given by Wei and Tanner (1990) to solve this integration problem. To do this, we need to generate a large number M of samples u_i from $f(u_i|y_i, z_i, w_i; \Gamma^{(t)})$. We know that

$$f(u_i|y_i, w_i, z_i; \Gamma^{(t)}) \propto f(y_i|u_i, z_i; \Gamma^{(t)}) f(u_i|w_i, z_i; \Gamma^{(t)}). \tag{4.11}$$

Note u_i is a $q_i \times 1$ vector. If $q > 1$ we can use the Gibbs sampler in appendix 4 to convert the multivariate distribution sampling problem $f(u_{i1}, \cdots, u_{iq}|y_i, w_i, z_i; \Gamma^{(t)})$ to a univariate conditional distribution problem $f(u_{i1}|u_{i2}, \cdots, u_{iq}, y_i, w_i, z_i; \Gamma^{(t)}), \cdots,$ $f(u_{iq}|u_{i1}, \cdots, u_{i(q-1)}, y_i, w_i, z_i; \Gamma^{(t)})$ first. Then, at the tth iteration, for each subject i, we interactively generate $u_i^{(k)}, k = 1, \cdots, M$ by Gibbs sampler along with the rejection sampling method based on (4.11). After that, choose $\psi^{(t+1)}$ and $\tau^{(t+1)}, \Sigma^{(t+1)}$ to maximize $\sum_{i=1}^{n} 1/M \sum_{k=1}^{M} logf(y_i|u_i^{(k)}, z_i; \psi^{(t)})$ and $\sum_{i=1}^{n} 1/M \sum_{k=1}^{M} logf(u_i^{(k)}|w_i, z_i; \tau^{(t)}, \Sigma^{(t)})$ respectively. If convergence is obtained, we can say $\psi^{(t+1)}$, $\tau^{(t+1)}$ and $\Sigma^{(t+1)}$ are the maximum likelihood estimation of parameters. Denote this by $\hat{\Gamma} = (\psi^{(t+1)}, \tau^{(t+1)}, \Sigma^{(t+1)})$.

The observed information matrix of the estimates $\hat{\Gamma}$ is

$$H = -Q'' = \ -\frac{1}{M} \sum_{i=1}^{n} \sum_{k=1}^{M} \frac{\partial^2}{\partial \Gamma \partial \Gamma'} l_i(\Gamma; y_i, u_i^{(k)}, z_i|\hat{\Gamma}). \tag{4.12}$$

4.2.2.3 Estimation of the parameters with covariate measurement error and missing information in the responses

As in Ibrahim et al. (2001) the missingness in the reponses can be expressed as

$$y_i = \begin{cases} y_{o,i} & \text{if } y_i \text{ is observed,} \\ y_{m,i} & \text{if } y_i \text{ is missing.} \end{cases} \tag{4.13}$$

and

$$r_i = \begin{cases} 0 & \text{if } y_i \text{ is observed,} \\ 1 & \text{if } y_i \text{ is missing.} \end{cases} \tag{4.14}$$

We suppose missing data indicator r_i follows independent Bernoulli distribution

$$f(r_i|y_i, x_i; \alpha) = (p_i)^{r_i}(1-p_i)^{1-r_i}, \tag{4.15}$$

where $p_i = P(r_i = 1)$. To connect the distribution of r_i to covariates the logistic regression is employed,

$$log\big[\frac{P(r_i=1)}{1-P(r_i=1)}\big] = V_i^T\alpha, \tag{4.16}$$

where V_i^T includes both missing response data, error prone covariates u_i and error free covariate z_i information, where α is the vector of parameters of missing data process.

The full joint likelihood considering covariate measurement error and missing response is

$$\prod_{i=1}^{n} f(y_i|u_i, z_i; \psi) f(r_i|u_i, z_i, y_i; \alpha) f(u_i|w_i, z_i; \tau, \Sigma), \tag{4.17}$$

where y_i is composed of observed part $y_{o,i}$ and missing part $y_{m,i}$. The complete

data log likelihood contributed from subject i is

$$\begin{aligned} l_i = \;& logf(y_i|u_i, z_i; \psi) + logf(r_i|u_i, z_i, y_i; \alpha) + logf(u_i|w_i, z_i; \tau, \Sigma) \\ = \;& -\log(1+\gamma) + \log[\gamma + f(0; \pi_i, \phi, \omega)]I_{\{y_i=0\}} + \log f(y_i; \pi_i, \phi, \omega)I_{\{y_i>0\}} \\ & + r_i * V_i^T\alpha - log(1 + e^{V_i^T\alpha}) + logf(u_i|w_i, z_i; \tau, \Sigma). \end{aligned} \tag{4.18}$$

Denote $\Gamma = (\psi, \tau, \Sigma, \alpha)$. The E-step provides the conditional expectation of the complete data log-likelihood with respect to the distribution of $y_{m,i}$ and u_i given the observed data and the current estimates of the parameters. Let t be an arbitrary number of iterations during maximization of the log-likelihood. Then given the observed data $(y_{o,i}, w_i, z_i, r_i)$ and current estimates of the parameters $\Gamma^{(t)}$, the conditional expectation of the complete data log-likelihood for the i^{th} observation in the $(s+1)^{th}$ iteration can be written as

$$\begin{aligned} Q_i(\Gamma|\Gamma^{(t)}) = \;& E\left[l_i|y_{o,i}, w_i, z_i, r_i; \Gamma^{(t)}\right] \\ = \;& \int\int \Big[log(1+\gamma) + \log[\gamma + f(0; \pi_i, \phi, \omega)]I_{\{y_i=0\}} \\ & + \log f(y_i; \pi_i, \phi, \omega)I_{\{y_i>0\}} \\ & + r_i * V_i^T\alpha - log(1 + e^{V_i^T\alpha}) + logf(u_i|w_i, z_i; \tau, \Sigma)\Big] \\ & \cdot f(y_{m,i}, u_i|y_{o,i}, r_i, w_i, z_i; \Gamma^{(t)})dy_{m,i}du_i. \end{aligned} \tag{4.19}$$

For each subject i, at the tth iteration, the kth $(k = 1, 2, ..., M)$ sample can be generated for $(y_{m,i}^{(k)}, u_i^{(k)})$ though Gibbs sampler along with the rejection sampling method in appendix 4 based on the following

$$\begin{aligned} f(y_{m,i}|y_{o,i}, u_i, w_i, z_i; \Gamma^{(t)}) \;& \propto f(y_i|u_i, z_i; \Gamma^{(t)})f(r_i|y_i, u_i, z_i; \Gamma^{(t)}), \\ f(u_i|y_i, w_i, z_i, r_i; \Gamma^{(t)}) \;& \propto f(y_i|u_i, z_i; \Gamma^{(t)}))f(u_i|w_i, z_i; \Gamma^{(t)})f(r_i|u_i, z_i, y_i; \Gamma^{(t)}). \end{aligned} \tag{4.20}$$

After replacing $(y_{i,m}, u_i)$ with $(y_{i,m}^{(k)}, u_i^{(k)})$, in the M step, we choose $\Gamma^{(t+1)}$ to maximize $\sum_{i=1}^{n} 1/M \sum_{k=1}^{M} Q_i$. If convergence is obtained, we can say $\Gamma^{(t+1)}$ is the maximum likelihood estimate of parameters. Denote this by $\hat{\Gamma} = \Gamma^{(t+1)}$.

The observed information matrix of the estimates $\hat{\Gamma}$ is

$$H = -Q'' = \ -\frac{1}{M}\sum_{i=1}^{n}\sum_{k=1}^{M}\frac{\partial^2}{\partial\Gamma\partial\Gamma'}l_i(\Gamma|\hat{\Gamma}). \tag{4.21}$$

4.3 Simulation study

A simulation study was conducted to investigate the properties of the estimates in terms of average of estimates (AE), relative bias (RB), variance (VAR), mean squared error (MSE) and coverage probability (CP) of estimates. The AE, RB, SE, MSE and CP, for example of $\hat{\pi}$, are obtained as: AE $(\hat{\pi}) = \frac{1}{N}\sum_{q=1}^{N}\hat{\pi}_q$, RB$(\hat{\pi})$= (AE -$\pi$)/$\pi$, VAR$(\hat{\pi}) = \frac{1}{N}\sum_{q=1}^{N}\widehat{var}(\hat{\pi}_q)$, where $\widehat{var}(\hat{\pi}_q)$ was obtained from the observed information matrix given in (4.12) or (4.21), MSE$(\hat{\pi}) = \frac{1}{N}\sum_{q=1}^{N}(\hat{\pi}_q - \pi)^2$, and CP$(\hat{\pi}) = \frac{1}{N}\sum_{q=1}^{N} I(\hat{\pi}_q - Z_{\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\pi}_q)} < \pi < \hat{\pi}_q + Z_{\frac{\alpha}{2}}\sqrt{\widehat{var}(\hat{\pi}_q)})$, where N is the number of samples we simulated.

4.3.1 Covariate measurement errors

For the case with one error prone covariate and one error free covariate we take $\pi_i = \exp(\beta_0 + \beta_1 u_i + \beta_2 z_i)/(1 + \exp(\beta_0 + \beta_1 u_i + \beta_2 z_i))$ with $\beta_0 = -1$, $\beta_1 = 1$ and $\beta_2 = 2$. Note that β_0 is the intercept parameter. Error free covariate z_i is generated from $N(0, 1)$.

We use Berkson measurement error model here. The surrogate variable w_i is generated from $N(1, 1)$. Then the true covariate can be generated from model $u_i = w_i + e_i$, where measurement error $e_i's$ are independent and identically following $N(0, \sigma^2)$.

In figure 4.1. we first illustrate the measurement error σ effect on the RB of parameter estimates if we use the observed data (y_i, w_i, z_i) directly without consid-

ering measurement error. We call this method the Naive method. Take $n = 100$, $m = 10$ and set different degree of measurement error (σ from 0.1 to 1). 2000 runs are performed. Apparently, if we ignore the measurement error the RB of parameter estimates will increase as the measurement error become larger. The measurement errors not only have an impact on the coefficient β of the error prone covariate but also on all the parameter estimations. When the measurement error increases, the over-dispersion parameter ϕ is affected a lot.

Now we compare the performance of the parameter of estimation between our proposed EM method with Naive method under three different degrees of measurement error ($\sigma = 0, \sigma = 0.5, \sigma = 0.9$) in Table 4.1 to Table 4.3. For empirical coverage probability we take nominal level $\alpha = 0.05$.

Results in Table 4.1 to Table 4.3 for ZIBB data without covariate measurement error show that all the parameters are well estimated irrespective of the sample sizes and at $n = 100$ show almost no estimation error. However, the coverage probability falls short of the nominal coverage of 95%.

When there is a covariate subjected to measurement error, the Naive method yields considerably larger AE, RB, SE and MSE and lower coverage probability, even for large sample size. The parameter π shows overestimation, whereas, β_0, β_1 and β_2 show underestimation. The parameter ϕ shows high RB (as high as 77%) for sample size($n = 100$) and measurement error ($\sigma = 0.9$).

The MCEM method, however, shows excellent performance in terms of all the measures for all five parameter estimates, except that the coverage probability for the parameter ϕ is shorter (ranges from .85 to .90) in comparison to that data without covariate measurement error. However, these coverage probabilities are much closer to the nominal coverage probability than those using the Naive method.

The RB, SE and MSE from Naive method tend to increase and the coverage probability tends to decrease as the degree of measurement error increases, However our proposed method shows good behaviors for all the estimates of the parameters although the measurement error goes up to $\sigma = 0.9$.

4.3.2 Covariate measurement errors and missing responses

For the case with one error prone covariate and one error free covariate with missing responses we also take $\pi_i = \exp(\beta_0 + \beta_1 u_i + \beta_2 z_i)/(1 + \exp(\beta_0 + \beta_1 u_i + \beta_2 z_i))$ with $\beta_0 = -1$, $\beta_1 = 1$ and $\beta_2 = 2$. Note that β_0 is the intercept parameter. The error free covariate z_i is generated from $N(0, 1)$.

We use Berkson measurement error model here. The surrogate variable w_i is generated from $N(1, 1)$. Then the true covariate can be generated from model $u_i = w_i + e_i$, where measurement errors $e_i's$ are independent and identically following $N(0, \sigma^2)$.

For the missing data process, we consider the logistic model

$$logit(P(r_i = 1)) = \alpha_0 + \alpha_1 x_i + \alpha_2 y_i, \tag{4.22}$$

from which missing data indicators r_i's are independently generated. The value of α_0 is set as 1.1 to make the baseline missing rate about 25%. The values of (α_1, α_2) are set as $(0, 0), (0.1, 0), (-0.1, 0.1)$ to indicate different missing data mechanisms.

We can see from model (4.22) that, when $\alpha_1 = 0$ and $\alpha_2 = 0$, the missing data do not depend on either the error prone covariate x_i or the missing response y_i, which results in MCAR. When $\alpha_1 \neq 0$ and $\alpha_2 = 0$, the missingness only depends on the error prone covariate x_i resulting in MAR. When $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$, the missingness depends on the missing response y_i, in addition to the error prone covariate x_i indicating that we have MNAR. Here, in order to control the missing rate close to the

baseline missing rate, we set small values for α_1 and α_2.

We compare the performance of parameter of estimation between our proposed MCEM method with the Naive method which only uses observed data directly without considering covariate measurement error and missing response under different degree of covariate measurement error ($\sigma = 0.5, \sigma = 0.9$) with three missing data mechanism in Table 4.4 to Table 4.12. We have discussed the performance of parameter of estimation for ZIBB model with missing response and without covariate measurement error ($\sigma = 0$) in Chapter 3. For empirical coverage probability we take nominal level $\alpha = 0.05$ here.

We can see from Table 4.4 to Table 4.12 that the proposed EM method works better than the Naive method for RB, VAR and MSE. In the presence of missing responses and covariate measurement error, the performance of the Naive method is affected remarkably, especially when the missingness probability depends on the covariate with measurement error and response, while the MCEM method performs steadily.

4.4 An Example: Analysis of a mutagenic data set

In this section we analyze a set of mutagenic data. The data obtained from Lüning et al. (1966) involved groups of male mice originating from an inbred CBA strain mated with groups of female mice originating from same inbred CBA strain. The experiment was conducted in three groups in which male mice were given 0 R, 300 R and 600 R respectively and then were mated within the first 7 days after irradiation.

The data are given in Table 4.13, and have been grouped according to the number of implants and the number of dead foetuses. We are interested in the dosage effect on

the death rate of the foetuses. The outcome variable is the number of dead foetuses in the litter. The independent variable is the dosage.

Since the data assigned the same exposure dose for each group, but the real exposure dose is particular to an individual, the exposure dose can be treated as an error prone covariate u_i. We employ the Berkson measurement error model here. We also control the litter size m_i as our error free covariate z_i. Then we fit the data using the zero-inflated beta-binomial model (4.3) with $\pi_i = \exp(\beta_0 + \beta_1 u_i + \beta_2 z_i)/(1 + \exp(\beta_0 + \beta_1 u_i + \beta_2 z_i))$, $u_i = treatment_i = 0, 300, 600$. where π_i is the proportion of dead implants, β_0 represents the intercept parameter, β_1 represents the regression parameter of treatment effect, and β_2 represents the regression parameter of litter size effect. Since the dosages u_i are far apart we standardize as $v_i = (u_i - \bar{u})/s$, where $\bar{u}$ and s are mean and standard deviation of the u_i values. The model then for the zero-inflated beta-binomial proportion becomes $\pi_i = \exp(\beta_0 + \beta_1 v_i + \beta_2 z_i)/(1 + \exp(\beta_0 + \beta_1 v_i + \beta_2 z_i))$. The maximum likelihood estimate (mle) of β_0, β_1, β_2, ϕ and ω for the mutagenic data are reported in Table 4.14. Both Naive and EM analysis methods suggest a positive dose effect and a negative litter size effect, however the magnitudes of the effect are very different. The EM method shows more dose effect than that revealed from the Naive method, but the estimate for liter size effect obtained from the Naive method is more than that calculated from the EM method. Moreover, the estimates of over-dispersion parameter ϕ and zero-inflation parameter ω from the Naive method are higher than those obtained from the EM method. The estimate of the measurement error parameter σ is 0.197 and its 95% confidence interval is $(0.191, 0.203)$ which indicates that measurement error exists in the exposure dose rate.

The data in Table 4.13 does not contain any missing values. However, in prac-

tice, in Toxicology and mutagenic studies, missingness can occur in addition to the covariate measurement error. So, to illustrate our method of analyzing mutagenic or toxicological data in the form of proportions that follow the ZIBB model, but contain missing responses in addition to covariate measurement error we generate missingness using the model (4.22). Estimates of the parameters β_0, β_1, and ϕ and their variances are given in Tables 4.15 to Table 4.17 for MCAR, MAR and MNAR respectively. The estimate of the measurement error parameter σ is 0.167, 0.158 and 0.105 for MCAR, MAR and MNAR respectively and their corresponding 95% confidence intervals are $(0.161, 0.204)$, $(0.121, 0.189)$ and $(0.102, 0.157)$ which indicate that measurement error exists in the exposure dose rate with all three missing mutagenic data.

4.5 Discussion

In this chapter, we have developed an estimation procedure for the parameters of a zero-inflated beta-binomial model in presence of covariate measurement error with or without missing response. We proposed the EM method to deal with covariate measurement error and missing response. The simulation studies for different degrees of covariate measurement error and different missing data mechanisms show that estimation by using the EM method performs well. Although the measurement error model and missing data mechanism have been discussed extensively in many articles, the current development for the estimation of the parameters of ZIBB in presence of covariate measurement error with missing response is new.

Moreover, we focus on structural modelling here by specifying a normally distributed covariate measurement error model for ZIBB data. In the measurement error literature, an alternative method is called functional modelling. Structural mod-

elling assumes a known distribution for the unobserved covariate U while functional modelling does not assume any distribution for U which is more robust.

We will develop a functional approach for ZIBB data with covariate measurement error by using the SIMEX (Carroll et al. (2006)) method in the future.

Figure 4.1: *Effect of different degrees of measurement error σ on the RB of estimates of the parameters under ZIBB model ignoring the covariate measurement error.*

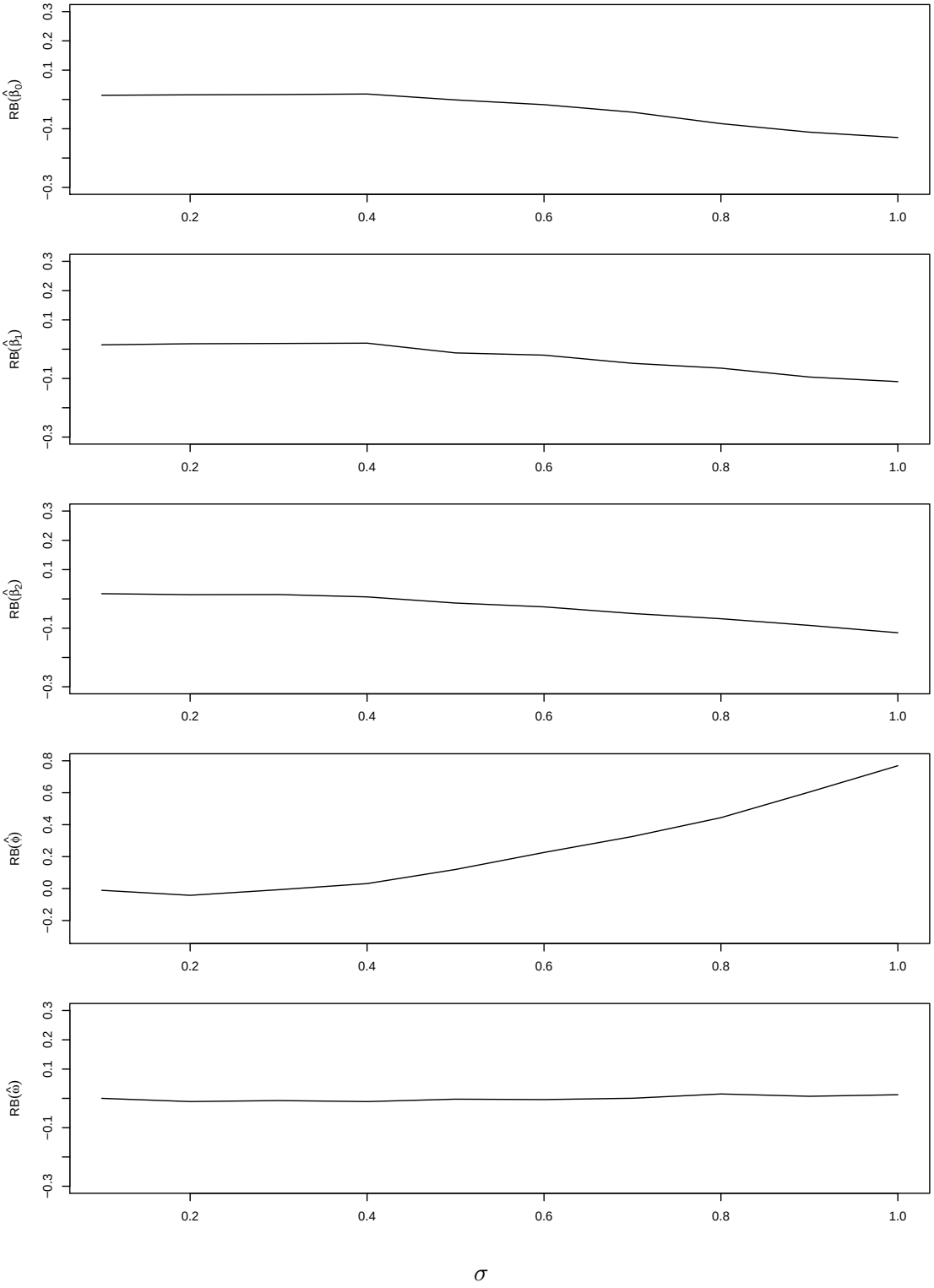


Table 4.1: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error, data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 30$).*

Method	Quantity	$\sigma = 0$					$\sigma = 0.5$					$\sigma = 0.9$				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=30	AE	-1.019	1.031	2.038	0.188	0.207	-0.960	0.961	1.938	0.241	0.210	-0.886	0.914	1.849	0.307	0.211
Naive	RB	0.019	0.031	0.019	-0.062	0.036	-0.040	-0.039	-0.031	0.206	0.051	-0.114	-0.086	-0.076	0.537	0.056
	VAR	0.161	0.080	0.160	0.011	0.007	0.241	0.132	0.210	0.021	0.009	0.263	0.140	0.219	0.033	0.010
	MSE	0.161	0.080	0.161	0.011	0.007	0.242	0.134	0.214	0.023	0.009	0.276	0.147	0.242	0.044	0.010
	CP	0.926	0.952	0.956	0.870	0.953	0.855	0.842	0.908	0.796	0.936	0.835	0.807	0.893	0.732	0.936
EM	AE	-1.019	1.031	2.038	0.188	0.207	-1.023	1.033	2.058	0.178	0.208	-0.963	0.983	1.981	0.233	0.210
	RB	0.019	0.031	0.019	-0.062	0.036	0.023	0.033	0.029	-0.112	0.041	-0.037	-0.017	-0.010	0.165	0.050
	VAR	0.161	0.080	0.160	0.011	0.007	0.186	0.091	0.202	0.013	0.009	0.204	0.097	0.209	0.020	0.009
	MSE	0.161	0.080	0.161	0.011	0.007	0.187	0.092	0.205	0.014	0.009	0.206	0.097	0.209	0.022	0.009
	CP	0.926	0.952	0.956	0.870	0.953	0.916	0.922	0.926	0.866	0.943	0.903	0.910	0.906	0.860	0.937

Table 4.2: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error, data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 50$).*

Method	Quantity	$\sigma = 0$					$\sigma = 0.5$					$\sigma = 0.9$				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=50	AE	-1.014	1.023	2.035	0.189	0.199	-0.974	0.966	1.953	0.250	0.201	-0.909	0.904	1.826	0.342	0.202
Naive	RB	0.014	0.023	0.018	-0.062	-0.005	-0.026	-0.034	-0.023	0.250	0.007	-0.091	-0.096	-0.087	0.709	0.009
	VAR	0.138	0.049	0.111	0.005	0.004	0.140	0.075	0.121	0.013	0.006	0.155	0.080	0.126	0.023	0.006
	MSE	0.138	0.049	0.112	0.005	0.004	0.141	0.076	0.123	0.016	0.006	0.163	0.089	0.156	0.043	0.006
	CP	0.931	0.936	0.940	0.881	0.958	0.835	0.829	0.893	0.787	0.934	0.857	0.829	0.895	0.689	0.929
EM	AE	-1.014	1.023	2.035	0.188	0.199	-1.024	1.033	2.042	0.182	0.199	-0.987	0.969	1.957	0.268	0.200
	RB	0.014	0.023	0.018	-0.062	-0.005	0.024	0.033	0.021	-0.092	-0.005	-0.013	-0.031	-0.022	0.338	0.000
	VAR	0.138	0.049	0.111	0.005	0.004	0.109	0.053	0.116	0.008	0.005	0.125	0.058	0.124	0.015	0.006
	MSE	0.138	0.049	0.112	0.005	0.004	0.110	0.054	0.118	0.008	0.005	0.125	0.059	0.126	0.020	0.006
	CP	0.931	0.936	0.940	0.881	0.958	0.921	0.926	0.930	0.875	0.953	0.922	0.904	0.909	0.870	0.944

Table 4.3: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error, data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 100$).*

Method	Quantity	$\sigma = 0$					$\sigma = 0.5$					$\sigma = 0.9$				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=100	AE	-0.999	1.000	2.017	0.198	0.200	-0.933	0.945	1.905	0.266	0.202	-0.872	0.880	1.757	0.354	0.205
Naive	RB	-0.001	0.000	0.009	-0.010	-0.001	-0.067	-0.055	-0.047	0.328	0.010	-0.128	-0.120	-0.122	0.772	0.024
	VAR	0.045	0.020	0.041	0.003	0.002	0.067	0.035	0.058	0.007	0.003	0.075	0.038	0.058	0.012	0.003
	MSE	0.045	0.020	0.041	0.003	0.002	0.071	0.038	0.067	0.011	0.003	0.091	0.053	0.118	0.036	0.003
	CP	0.936	0.949	0.941	0.902	0.957	0.875	0.850	0.923	0.699	0.917	0.852	0.805	0.848	0.655	0.922
EM	AE	-0.999	1.000	2.017	0.198	0.200	-0.997	1.000	2.037	0.197	0.200	-0.931	0.932	1.873	0.284	0.203
	RB	-0.001	0.000	0.009	-0.010	-0.001	-0.003	0.000	0.019	-0.015	-0.001	-0.069	-0.068	-0.064	0.422	0.013
	VAR	0.045	0.020	0.041	0.003	0.002	0.054	0.026	0.057	0.004	0.003	0.061	0.028	0.058	0.008	0.003
	MSE	0.045	0.020	0.041	0.003	0.002	0.054	0.026	0.058	0.004	0.003	0.066	0.033	0.074	0.015	0.003
	CP	0.936	0.949	0.941	0.902	0.957	0.926	0.939	0.931	0.890	0.957	0.904	0.905	0.912	0.875	0.944

Table 4.4: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error and the missness does not depend on covarites and response variable (MCAR), data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 30$).*

Method	Quantity	$\sigma = 0$(complete data)					$\sigma = 0.5$(baseline missing rate 25%)					$\sigma = 0.9$(baseline missing rate 25%)				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=30	AE	-1.019	1.031	2.038	0.188	0.207	-0.891	0.903	1.810	0.238	0.192	-0.816	0.892	1.880	0.303	0.190
	RB	0.019	0.031	0.019	-0.062	0.036	-0.109	-0.097	-0.095	0.190	-0.040	-0.184	-0.108	-0.060	0.515	-0.050
Naïve	VAR	0.161	0.080	0.160	0.011	0.007	0.286	0.185	0.223	0.065	0.032	0.292	0.191	0.253	0.085	0.065
	MSE	0.161	0.080	0.161	0.011	0.007	0.298	0.194	0.259	0.066	0.032	0.326	0.203	0.267	0.096	0.065
	CP	0.926	0.952	0.956	0.870	0.953	0.801	0.752	0.802	0.798	0.832	0.713	0.705	0.753	0.748	0.805
	AE	-1.019	1.031	2.038	0.188	0.207	-0.987	0.978	2.020	0.198	0.198	0.935	0.971	1.956	0.204	0.195
	RB	0.019	0.031	0.019	-0.062	0.036	-0.013	-0.022	0.010	-0.010	-0.010	-1.935	-0.029	-0.022	0.020	-0.025
EM	VAR	0.161	0.080	0.160	0.011	0.007	0.256	0.168	0.225	0.045	0.028	0.276	0.192	0.253	0.069	0.068
	MSE	0.161	0.080	0.161	0.011	0.007	0.256	0.168	0.225	0.045	0.028	4.020	0.193	0.255	0.069	0.068
	CP	0.926	0.952	0.956	0.870	0.953	0.938	0.945	0.941	0.908	0.912	0.932	0.935	0.940	0.891	0.907

Table 4.5: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error and the missness does not depend on covarites and response variable (MCAR), data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 50$).*

Method	Quantity	$\sigma = 0$(complete data)					$\sigma = 0.5$(baseline missing rate 25%)					$\sigma = 0.9$(baseline missing rate 25%)				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=50	AE	-1.014	1.023	2.035	0.189	0.199	-0.970	0.985	1.971	0.240	0.193	-0.865	0.884	1.886	0.338	0.192
	RB	0.014	0.023	0.018	-0.062	-0.005	-0.030	-0.015	-0.015	0.200	-0.035	-0.135	-0.116	-0.057	0.690	-0.040
Naïve	VAR	0.138	0.049	0.111	0.005	0.004	0.198	0.165	0.227	0.022	0.030	0.231	0.185	0.275	0.062	0.061
	MSE	0.138	0.049	0.112	0.005	0.004	0.199	0.165	0.228	0.024	0.030	0.249	0.198	0.288	0.081	0.061
	CP	0.931	0.936	0.940	0.881	0.958	0.801	0.751	0.762	0.796	0.858	0.752	0.653	0.735	0.723	0.801
	AE	-1.014	1.023	2.035	0.189	0.199	-0.981	0.985	1.986	0.198	0.202	-0.955	0.978	2.025	0.196	0.198
	RB	0.014	0.023	0.018	-0.062	-0.005	-0.019	-0.015	-0.007	-0.010	0.010	-0.045	-0.022	0.013	-0.020	-0.010
EM	VAR	0.138	0.049	0.111	0.005	0.004	0.182	0.145	0.165	0.054	0.025	0.198	0.162	0.202	0.063	0.068
	MSE	0.138	0.049	0.112	0.005	0.004	0.182	0.145	0.165	0.054	0.025	0.200	0.162	0.203	0.063	0.068
	CP	0.931	0.936	0.940	0.881	0.958	0.940	0.951	0.952	0.913	0.922	0.934	0.945	0.941	0.910	0.917

Table 4.6: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error and the missness does not depend on covarites and response variable (MCAR), data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 100$).*

Method	Quantity	$\sigma = 0$(complete data)					$\sigma = 0.5$(baseline missing rate 25%)					$\sigma = 0.9$(baseline missing rate 25%)				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=100	AE	-0.999	1.000	2.017	0.198	0.200	-0.951	0.946	1.835	0.259	0.199	-0.877	0.891	1.901	0.349	0.203
	RB	-0.001	0.000	0.009	-0.010	-0.001	-0.049	-0.054	-0.083	0.295	-0.005	-0.123	-0.109	-0.050	0.745	0.015
Naïve	VAR	0.045	0.020	0.041	0.003	0.002	0.079	0.032	0.058	0.031	0.020	0.200	0.145	0.198	0.054	0.058
	MSE	0.045	0.020	0.041	0.003	0.002	0.081	0.035	0.085	0.034	0.020	0.215	0.157	0.208	0.076	0.058
	CP	0.936	0.949	0.941	0.902	0.957	0.891	0.798	0.805	0.756	0.901	0.851	0.782	0.756	0.690	0.825
	AE	-0.999	1.000	2.017	0.198	0.200	-0.986	0.990	1.988	0.197	0.199	-0.986	0.952	1.945	0.198	0.198
	RB	-0.001	0.000	0.009	-0.010	-0.001	-0.014	-0.010	-0.006	-0.015	-0.005	-0.014	-0.048	-0.028	-0.010	-0.010
EM	VAR	0.045	0.020	0.041	0.003	0.002	0.071	0.038	0.065	0.029	0.018	0.161	0.084	0.098	0.010	0.045
	MSE	0.045	0.020	0.041	0.003	0.002	0.071	0.038	0.065	0.029	0.018	0.161	0.086	0.101	0.010	0.045
	CP	0.936	0.949	0.941	0.902	0.957	0.934	0.949	0.954	0.901	0.935	0.949	0.950	0.943	0.910	0.933

Table 4.7: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error and the missingness only depends on error prone covarite(MAR), data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 30$).*

Method	Quantity	$\sigma = 0$(complete data)					$\sigma = 0.5$(baseline missing rate 25%)					$\sigma = 0.9$(baseline missing rate 25%)				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=30	AE	-1.019	1.031	2.038	0.188	0.207	-0.878	0.892	1.801	0.241	0.184	-0.806	0.881	1.780	0.310	0.182
	RB	0.019	0.031	0.019	-0.062	0.036	-0.122	-0.108	-0.100	0.205	-0.078	-0.194	-0.119	-0.110	0.550	-0.090
Naïve	VAR	0.161	0.080	0.160	0.011	0.007	0.298	0.198	0.256	0.038	0.010	0.301	0.206	0.301	0.099	0.097
	MSE	0.161	0.080	0.161	0.011	0.007	0.313	0.210	0.296	0.040	0.010	0.339	0.220	0.349	0.111	0.097
	CP	0.926	0.952	0.956	0.870	0.953	0.721	0.701	0.752	0.791	0.820	0.613	0.635	0.672	0.698	0.723
	AE	-1.019	1.031	2.038	0.188	0.207	-0.945	0.964	1.892	0.205	0.196	-0.902	0.951	1.869	0.208	0.195
	RB	0.019	0.031	0.019	-0.062	0.036	-0.055	-0.036	-0.054	0.025	-0.020	-0.098	-0.049	-0.066	0.040	-0.025
EM	VAR	0.161	0.080	0.160	0.011	0.007	0.268	0.179	0.234	0.075	0.077	0.298	0.196	0.288	0.085	0.084
	MSE	0.161	0.080	0.161	0.011	0.007	0.271	0.180	0.246	0.035	0.009	0.308	0.198	0.305	0.085	0.084
	CP	0.926	0.952	0.956	0.870	0.953	0.932	0.941	0.931	0.876	0.882	0.911	0.922	0.925	0.861	0.884

Table 4.8: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error and the missingness only depends on error prone covarite(MAR), data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 50$).*

Method	Quantity	$\sigma = 0$(complete data)					$\sigma = 0.5$(baseline missing rate 25%)					$\sigma = 0.9$(baseline missing rate 25%)				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=50	AE	-1.014	1.023	2.035	0.189	0.199	-0.892	0.903	1.915	0.254	0.187	-0.858	0.876	1.853	0.341	0.185
	RB	0.014	0.023	0.018	-0.062	-0.005	-0.108	-0.097	-0.043	0.270	-0.065	-0.142	-0.124	-0.074	0.705	-0.075
Naïve	VAR	0.161	0.080	0.160	0.011	0.007	0.206	0.174	0.235	0.026	0.067	0.254	0.192	0.296	0.084	0.083
	MSE	0.138	0.049	0.112	0.005	0.004	0.218	0.183	0.242	0.029	0.067	0.274	0.207	0.318	0.104	0.083
	CP	0.931	0.936	0.940	0.881	0.958	0.752	0.691	0.712	0.725	0.821	0.612	0.515	0.618	0.689	0.713
	AE	-1.014	1.023	2.035	0.189	0.199	-0.975	0.978	1.966	0.198	0.198	-0.970	0.973	1.953	0.197	0.194
	RB	0.014	0.023	0.018	-0.062	-0.005	-0.025	-0.022	-0.017	-0.010	-0.010	-0.030	-0.027	-0.024	-0.015	-0.030
EM	VAR	0.161	0.080	0.160	0.011	0.007	0.196	0.154	0.202	0.067	0.053	0.234	0.176	0.254	0.075	0.076
	MSE	0.138	0.049	0.112	0.005	0.004	0.197	0.154	0.203	0.024	0.006	0.235	0.177	0.256	0.075	0.076
	CP	0.931	0.936	0.940	0.881	0.958	0.931	0.941	0.942	0.905	0.903	0.920	0.932	0.934	0.886	0.891

Table 4.9: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error and the missingness only depends on error prone covarite(MCR), data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 100$).*

Method	Quantity	$\sigma = 0$(complete data)					$\sigma = 0.5$(baseline missing rate 25%)					$\sigma = 0.9$(baseline missing rate 25%)				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=100	AE	-0.999	1.000	2.017	0.198	0.200	-0.901	0.905	1.902	0.265	0.192	-0.872	0.859	1.864	0.358	0.190
Naïve	RB	-0.001	0.000	0.009	-0.010	-0.001	-0.099	-0.095	-0.049	0.325	-0.040	-0.128	-0.141	-0.068	0.790	-0.050
	VAR	0.045	0.020	0.041	0.003	0.002	0.086	0.045	0.067	0.035	0.034	0.206	0.168	0.212	0.065	0.069
	MSE	0.045	0.020	0.041	0.003	0.002	0.096	0.054	0.077	0.039	0.034	0.222	0.188	0.230	0.090	0.069
	CP	0.936	0.949	0.941	0.902	0.957	0.821	0.732	0.789	0.701	0.896	0.801	0.752	0.706	0.653	0.791
EM	AE	-0.999	1.000	2.017	0.198	0.200	-0.981	0.986	1.971	0.195	0.197	-0.961	1.028	1.962	0.115	0.194
	RB	-0.001	0.000	0.009	-0.010	-0.001	-0.019	-0.014	-0.015	-0.025	-0.017	-0.039	0.028	-0.019	-0.423	-0.031
	VAR	0.045	0.020	0.041	0.003	0.002	0.081	0.042	0.068	0.032	0.028	0.172	0.095	0.123	0.012	0.058
	MSE	0.045	0.020	0.041	0.003	0.002	0.081	0.042	0.069	0.032	0.028	0.174	0.096	0.124	0.019	0.058
	CP	0.936	0.949	0.941	0.902	0.957	0.936	0.950	0.952	0.910	0.906	0.931	0.940	0.942	0.901	0.905

Table 4.10: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error and the missingness depends on error prone covarites and reponse variable (MNAR), data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 30$).*

Method	Quantity	$\sigma = 0$(complete data)					$\sigma = 0.5$(baseline missing rate 25%)					$\sigma = 0.9$(baseline missing rate 25%)				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=30	AE	-1.019	1.031	2.038	0.188	0.207	-0.830	0.896	1.872	0.298	0.213	-0.791	0.802	1.675	0.356	0.223
	RB	0.019	0.031	0.019	-0.062	0.036	-0.170	-0.104	-0.064	0.488	0.067	-0.209	-0.198	-0.163	0.780	0.115
Naïve	VAR	0.161	0.080	0.160	0.011	0.007	0.379	0.201	0.303	0.043	0.063	0.393	0.288	0.401	0.103	0.087
	MSE	0.161	0.080	0.161	0.011	0.007	0.408	0.211	0.319	0.052	0.063	0.436	0.327	0.507	0.128	0.088
	CP	0.926	0.952	0.956	0.870	0.953	0.712	0.689	0.705	0.724	0.814	0.601	0.521	0.601	0.651	0.701
	AE	-1.019	1.031	2.038	0.188	0.207	-0.914	0.966	1.954	0.184	0.207	-0.905	0.941	1.932	0.180	0.212
	RB	0.019	0.031	0.019	-0.062	0.036	-0.086	-0.034	-0.023	-0.079	0.036	-0.095	-0.059	-0.034	-0.100	0.060
EM	VAR	0.161	0.080	0.160	0.011	0.007	0.304	0.198	0.286	0.038	0.054	0.363	0.276	0.321	0.101	0.081
	MSE	0.161	0.080	0.161	0.011	0.007	0.311	0.199	0.288	0.038	0.054	0.372	0.279	0.326	0.101	0.081
	CP	0.926	0.952	0.956	0.870	0.953	0.924	0.931	0.926	0.862	0.872	0.901	0.913	0.912	0.831	0.851

Table 4.11: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error and the missingness depends on error prone covarites and reponse variable (MNAR), data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 50$).*

Method	Quantity	$\sigma = 0$(complete data)					$\sigma = 0.5$(baseline missing rate 25%)					$\sigma = 0.9$(baseline missing rate 25%)				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=50	AE	-1.014	1.023	2.035	0.189	0.199	-0.840	0.897	1.847	0.328	0.207	-0.805	0.856	1.712	0.330	0.219
Naïve	RB	0.014	0.023	0.018	-0.062	-0.005	-0.160	-0.103	-0.076	0.639	0.036	-0.195	-0.144	-0.144	0.650	0.095
	VAR	0.161	0.080	0.160	0.011	0.007	0.208	0.108	0.170	0.029	0.071	0.302	0.226	0.298	0.092	0.082
	MSE	0.138	0.049	0.112	0.005	0.004	0.234	0.119	0.193	0.045	0.071	0.340	0.247	0.381	0.109	0.082
	CP	0.931	0.936	0.940	0.881	0.958	0.752	0.691	0.712	0.691	0.821	0.612	0.515	0.618	0.689	0.713
EM	AE	-1.014	1.023	2.035	0.189	0.199	-0.916	0.974	2.089	0.208	0.196	-0.909	0.952	2.097	0.212	0.205
	RB	0.014	0.023	0.018	-0.062	-0.005	-0.084	-0.026	0.044	0.040	-0.020	-0.091	-0.048	0.049	0.060	0.025
	VAR	0.161	0.080	0.160	0.011	0.007	0.198	0.101	0.162	0.022	0.055	0.288	0.203	0.225	0.075	0.068
	MSE	0.138	0.049	0.112	0.005	0.004	0.205	0.102	0.170	0.022	0.055	0.296	0.205	0.234	0.075	0.068
	CP	0.931	0.936	0.940	0.881	0.958	0.926	0.932	0.931	0.886	0.882	0.901	0.920	0.921	0.879	0.865

Table 4.12: *Properties (AE, RB, VAR, MSE, CP) of the estimates of the parameters under different degrees of covariate measurement error and the missingness depends on error prone covarites and reponse variable (MNAR), data simulated from ZIBB(π, ϕ, ω), based on 1000 simulation runs ($n = 100$).*

Method	Quantity	$\sigma = 0$(complete data)					$\sigma = 0.5$(baseline missing rate 25%)					$\sigma = 0.9$(baseline missing rate 25%)				
		$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$	$\beta_0 = -1$	$\beta_1 = 1$	$\beta_2 = 2$	$\phi = 0.2$	$\omega = 0.2$
n=100	AE	-0.999	1.000	2.017	0.198	0.200	-0.864	0.890	1.791	0.346	0.205	-0.821	0.871	1.689	0.360	0.211
Naïve	RB	-0.001	0.000	0.009	-0.010	-0.001	-0.136	-0.110	-0.104	0.732	0.025	-0.179	-0.129	-0.156	0.800	0.055
	VAR	0.045	0.020	0.041	0.003	0.002	0.099	0.050	0.079	0.015	0.046	0.123	0.078	0.082	0.025	0.051
	MSE	0.045	0.020	0.041	0.003	0.002	0.117	0.062	0.122	0.036	0.046	0.155	0.095	0.179	0.051	0.051
	CP	0.936	0.949	0.941	0.902	0.957	0.801	0.721	0.752	0.662	0.832	0.786	0.715	0.689	0.557	0.721
EM	AE	-0.999	1.000	2.017	0.198	0.200	-0.927	0.965	2.048	0.200	0.198	-0.912	1.018	2.222	0.205	0.205
	RB	-0.001	0.000	0.009	-0.010	-0.001	-0.073	-0.035	0.024	0.001	-0.010	-0.088	0.018	0.111	0.025	0.025
	VAR	0.045	0.020	0.041	0.003	0.002	0.087	0.045	0.068	0.014	0.037	0.102	0.068	0.071	0.020	0.042
	MSE	0.045	0.020	0.041	0.003	0.002	0.092	0.046	0.070	0.014	0.037	0.110	0.068	0.120	0.020	0.042
	CP	0.936	0.949	0.941	0.902	0.957	0.934	0.945	0.941	0.901	0.905	0.904	0.912	0.915	0.889	0.886

Table 4.13: *The number of females with* $0, 1, 2,$ *etc. dead implants when 5-10 zygotes were implanted after matings during the first* 7 *days after irradiation of males with* 0 *(control), 300 R and 600 R.*

Dose		Dead Implants									
(R)	Implantations	0	1	2	3	4	5	6	7	Total	dead
0	5	30	27	9	5	_	_	_	_	71	16.9
	6	86	51	14	4	1	_	_	_	156	10.1
	7	111	73	31	8	1	_	_	_	224	10.4
	8	79	44	23	3	_	1	_	_	150	8.7
	9	32	29	8	1	_	_	_	_	70	7.6
	10	5	5	2	_	_	_	_	_	12	7.5
300	5	27	41	32	17	4	_	_	_	121	28.4
	6	28	47	59	28	6	1	1	_	170	27.8
	7	31	61	54	20	19	1	_	_	186	23.8
	8	12	32	24	22	8	1	_	_	99	23.1
	9	1	6	9	6	1	1	_	_	24	23.6
	10	1	2	1	_	_	_	_	_	4	10.0
600	5	16	32	48	49	15	_	_	_	160	41.9
	6	7	35	45	37	20	9	_	_	153	39.3
	7	5	22	27	36	17	9	3	1	120	38.3
	8	1	4	12	11	8	7	_	2	45	39.4
	9	_	_	2	2	2	_	1	_	7	38.1
	10	_	_	_	_	_	_	_	1	1	70.0

Table 4.14: *Estimates, standard error, variance and confidence interval of the parameters for mutagenic data.*

Method	Quantity	β_0	β_1	β_2	ϕ	ω
	MLE	-0.638	0.674	-0.098	0.012	0.011
	SE	0.097	0.024	0.013	0.007	0.011
Naive	VAR	0.009	0.001	0.000	0.000	0.000
	95% LB	-0.828	0.627	-0.123	-0.002	-0.011
	95% UB	-0.448	0.721	-0.073	0.026	0.033
	MLE	-0.754	0.731	-0.085	0.010	0.009
	SE	0.083	0.021	0.012	0.005	0.010
EM	VAR	0.007	0.000	0.000	0.000	0.000
	95% LB	-0.917	0.690	-0.109	0.000	-0.011
	95% UB	-0.591	0.772	-0.061	0.020	0.029

Table 4.15: *Estimates, standard error, variance and confidence interval of the parameters for mutagenic data under MCAR.*

Method	Quantity	β_0	β_1	β_2	ϕ	ω
	MLE	-0.752	0.682	-0.084	0.017	0.004
	SE	0.111	0.027	0.015	0.009	0.013
Naive	VAR	0.012	0.001	0.000	0.000	0.000
	95% LB	-0.970	0.629	-0.113	-0.001	-0.021
	95% UB	-0.534	0.735	-0.055	0.035	0.029
	MLE	-0.816	0.738	-0.080	0.014	0.002
	SE	0.102	0.024	0.012	0.007	0.011
EM	VAR	0.010	0.001	0.000	0.000	0.000
	95% LB	-1.016	0.691	-0.104	0.000	-0.020
	95% UB	-0.616	0.785	-0.056	0.028	0.024

Table 4.16: *Estimates, standard error, variance and confidence interval of the parameters for mutagenic data under MAR.*

Method	Quantity	β_0	β_1	β_2	ϕ	ω
	MLE	-0.690	0.676	-0.090	0.014	0.006
	SE	0.121	0.028	0.017	0.008	0.011
Naive	VAR	0.015	0.001	0.000	0.000	0.000
	95% LB	-0.927	0.621	-0.123	-0.002	-0.016
	95% UB	-0.453	0.731	-0.057	0.030	0.028
	MLE	-0.741	0.734	-0.075	0.010	0.004
	SE	0.115	0.021	0.012	0.006	0.001
EM	VAR	0.013	0.000	0.000	0.000	0.000
	95% LB	-0.966	0.693	-0.099	0.098	0.002
	95% UB	-0.516	0.775	-0.051	0.122	0.006

Table 4.17: *Estimates, standard error, variance and confidence interval of the parameters for mutagenic data under MNAR.*

Method	Quantity	β_0	β_1	β_2	ϕ	ω
	MLE	-1.111	0.982	-0.133	0.005	0.124
	SE	0.140	0.324	0.072	0.026	0.160
Naive	VAR	0.020	0.105	0.005	0.001	0.026
	95% LB	-1.385	0.347	-0.274	-0.046	-0.190
	95% UB	-0.837	1.617	0.008	0.056	0.438
	MLE	-0.812	0.667	-0.072	0.020	0.004
	SE	0.140	0.324	0.065	0.026	0.160
EM	VAR	0.020	0.105	0.005	0.001	0.026
	95% LB	-1.086	0.032	-0.213	-0.031	-0.310
	95% UB	-0.538	1.302	0.069	0.071	0.318

Chapter 5

Summary and Plan for Future Study

5.1 Summary

We have developed estimation procedures for the parameters of a zero-inflated beta-binomial model in presence of missing responses with or without covariate measurement error. We have applied a weighted expectation maximization algorithm for the maximum likelihood estimation of the parameters. Although missing data methodologies and measurement error procedure have been discussed extensively in the literature, the current development for the estimation of the parameters of zero-inflated beta-binomial model in presence of missing responses with/without covariate measurement error is new.

In chapter 2, we proposed an estimation procedure for the parameters of a zero-inflated beta-binomial model in presence of missing responses only. The general findings through simulations and data analyses are:

(a) Data without covariates: For complete data and under MCAR and MAR, all the parameters are well estimated irrespective of the sample sizes and percentage missing. All of the AE, RB, VAR, and MSE show good behavior. However, all the parameter estimates show shorter coverage probability, especially for ϕ, whose coverage probability ranges from 0.91 to 0.93. Under MNAR, the CC method for all the parameters yields considerably larger AE, RB, SE and MSE and lower coverage probability, even for large sample size. The EM method shows excellent performance in terms of all the measures for all three parameter estimates, except that the coverage probability for the parameter ϕ is shorter (ranges from .87 to .91) in comparison to that from complete data. However, these coverage probabilities are much closer to the nominal coverage probability than those using the CC method. All parameters are well estimated even at 25% baseline missing.

(b) Data with one covariate: Results for complete data are almost the same as those with no covariate except that to see such good behavior much larger sample sizes are required. Similarly, estimates of all the parameters, under MAR and MNAR, show similar behavior as those with no covariates except that it now requires much larger sample sizes.

In chapter 4, we have developed an estimation procedure for the parameters of a zero-inflated beta-binomial model in presence of covariate measurement error with or without missing response. We proposed the EM method for dealing with covariate measurement error with or without missing response. The simulation studies for different degrees of covariate measurement error and three missing data mechanisms show that estimation by using the EM method performs well in terms of the properties of the estimates using different measures, such as, average estimates (AE), relative bias (RB), variance(VAR), mean squared error (MSE) and coverage probability (CP)

of estimates.

5.2 Plan for Future Study: A Random Effects Transition Model For Longitudinal Binary Data With Missing Response And Covariate Measurement Error

In many biomedical studies, such as, the study of drug use, the probability of current binary response depends on a previous binary response. For example, the probability of a child having an obesity problem at time t_{ij} depends not only on explanatory variables, but also on the obesity status at time $t_{i(j-1)}$. A transition model is useful in such situations.

When we are interested in the dynamic features of transition patterns in repeated measurements, an appropriate longitudinal way is to model the transition probabilities over the study period. Longitudinal designs in bio-medical studies often collect data on binary repeated measures that indicate the presence or absence of clinical or biological states. Binary repeated measures can be conveniently modeled by Markov chains with transition probabilities, for example, the probability of changing from use to no use of a certain drug (or vice verse) in drug abuse treatment research. This strategy brings intuitive statistical interpretation to the study of dynamic changes in response to treatment through time and across subjects. Key targets of inference include the probability that subjects in a specific condition shift from use to non use and the probability that subjects maintain non use throughout the trial (Yang et

al. (2007)). For complete longitudinal data, Markov transition models have been studied by several authors. For example, Korn and Whittemore (1979) model the probability of occupying the current state using the previous state. Wu and Ware (1979) assume one binary event (e.g., death) though the covariate information as time passes before the event. Zeger and Qaqish (1988) discuss a class of Markov regression models for time-series by using a quasi-likelihood approach. Zeng and Cook (2007) propose a estimation method based on joint transition models for multivariate longitudinal binary data using GEE2. For incomplete data, Deltour et al. (1999) use stochastic algorithms for Markov models estimation with intermittent missing data. Albert (2000) develops a transitional model for longitudinal binary data, subject to nonignorable missing data and proposes an EM algorithm for parameter estimation. In Albert and Follmann (2003), an extended version of the Markov transition model was proposed to handle nonignorable missing values in a binary longitudinal data set.

Measurement error happens when there is a difference between a measured value of quantity and its true value. For example, measurable values are inconsistent when repeated measures of a constant attribute or quantity are taken. Errors can also be introduced by an inaccurate instrument(method) used in the experiment.

Let Y_{ij} be the outcome variable for subject i at the jth time point, $X_{ij} = (X_{ij1}, \cdots, X_{ijp})$ be the vector of p covariates, $i = 1, ..., n$, and $j = 1, ..., m_i$. Denote $Y_i = (Y_{i1}, ..., Y_{im_i})^T$ and $X_i = (X_{i1}, ..., X_{im_i})^T$. A transition model for binary response data is,

$$logit(P(Y_{ij} = 1|\mathbf{H}_{ij}) = X_{ij}\beta + \kappa(\mathbf{H}_{ij}, \alpha), \tag{5.1}$$

where $\mathbf{H}_{ij} = (Y_{i1}, Y_{i2}, ..., Y_{i(j-1)})$ and $\kappa(.)$ is a function of previous observations and the current observation.

A one-step Markov transition model assumes that $y_{ij}(j > 1)$ is independent of earlier observations given the previous observation $y_{i(j-1)}$. A simple transition model which assumes a first-order Markov process for the response can be written as

$$logit(P(Y_{ij} = 1|\mathbf{H}_{ij}) = X_{ij}\beta + Y_{i(j-1)}\alpha. \tag{5.2}$$

To capture the baseline heterogeneity across subjects (Albert and Follmann, 2003) a random intercept effect can be used

$$\left.\begin{aligned} logit(P_{01}(\xi_i)) &= logit[P(Y_{ij} = 1|Y_{i(j-1)} = 0, \xi_i)]) = X_{ij}\beta_{01} + \xi_i \\ logit(P_{10}(\xi_i)) &= logit[P(Y_{ij} = 0|Y_{i(j-1)} = 1, \xi_i)]) = X_{ij}\beta_{10} + \nu\xi_i \end{aligned}\right\}, \tag{5.3}$$

from which we obtain

$$\left.\begin{aligned} P_{01}(\xi_i) &= \tfrac{\exp(X_{ij}\beta_{01}+\xi_i)}{1+\exp(X_{ij}\beta_{01}+\xi_i)} \\ P_{10}(\xi_i) &= \tfrac{exp(X_{ij}\beta_{10}+\nu\xi_i)}{1+\exp(X_{ij}\beta_{10}+\nu\xi_i)} \end{aligned}\right\}, \tag{5.4}$$

where β_{01}, β_{10} are regression parameters, and ξ_i is the random effect distributed as $N(0, \sigma^2)$. See Albert and Follmann (2003). The parameter ν represents the association between $P_{01}(\xi_i)$ and $P_{10}(\xi_i)$. Note that $P_{01} + P_{00} = 1 = P_{11} + P_{10} = 1$. To see the effect of ν for some fixed parameters see Figure 3 (Albert and Follmann (2003)). Model (5.3) is the random effect transition model (see Albert and Follmann (2003) and Yang et al. (2007)).

Let $\theta = (\beta_{01}, \beta_{10})$. Thus, the model for y_{ij} given x_{ij}, $y_{i,j-1}$, ξ_i, and θ can be written as

$$f(y_{ij}|x_{ij}, z_{ij}, y_{i,j-1}, \xi_i, \theta) = \begin{cases} P_{01}^{y_{ij}}(\xi_i)(1 - P_{01}(\xi_i))^{1-y_{ij}} & \text{if } y_{i(j-1)} = 0 \\ P_{10}^{1-y_{ij}}(\xi_i)(1 - P_{10}(\xi_i))^{y_{ij}} & \text{if } y_{i(j-1)} = 1. \end{cases} \tag{5.5}$$

Transition models are most appropriate when interest lies in understanding how changes in the response occur over time and how covariates alter the governing transition probabilities (see Zeng and Cook (2007)). Models (5.5) can therefore be used

to analyze complete binary longitudinal data. However, in practice some subjects may not be available at all time points resulting in missing observations. Data analysis may be further complicated when measurement error occurs in some covariates. The purpose of this paper is to develop inference procedures for parameters of the random effects transition model (5.5) for longitudinal data having missing responses and covariate measurement error. Four scenarios are considered: (a) no missing data and no measurement error, (b) no missing data and measurement error (b) missing data and no measurement error, and (d) missing data and measurement error.

In this paper we only consider the missing response which only incorporates intermittent missing with missingness indicators defined in Ibrahim et al. (2001). The complete data and missingness can be expressed as

$$y_i = \begin{cases} y_{o,i} & \text{if } y_i \text{ is observed,} \\ y_{m,i} & \text{if } y_i \text{ is missing.} \end{cases} \tag{5.6}$$

and

$$r_i = \begin{cases} 0 & \text{if } y_i \text{ is observed,} \\ 1 & \text{if } y_i \text{ is missing.} \end{cases} \tag{5.7}$$

We suppose missing data indicator r_i follows

$$f(r_i|y_i, x_i; \alpha) = (p_i)^{r_i}(1-p_i)^{1-r_i}, \tag{5.8}$$

where $p_i = P(r_i = 1)$. To connect the distribution of r_i to covariates, logistic regression is employed,

$$log\big[\frac{P(r_i=1)}{1-P(r_i=1)}\big] = w_i^T\phi, \tag{5.9}$$

where w_i^T includes both missing data and observed data information, ϕ is the vector of parameters of the missing data process.

Data analysis may be further complicated when measurement error occurs in some covariates. Following Carroll et al. (2006) we model the error prone covariate x_{ij} as

$$x_{ij} = \gamma_0 + \gamma_1 u_{ij} + \gamma_2 z_{ij} + e_{ij},$$

where u_{ij} is the observed value for x_{ij}, z_{ij} is an error free covariate and e_{ij} is an error term which follows $N(0, \delta^2)$. We define $\omega = (\gamma_0, \gamma_1, \gamma_2,)$. We have the density function for x_{ij} as

$$f(x_{ij}|u_{ij}, z_{ij}, \omega) = \frac{1}{\sqrt{2\pi\delta^2}} exp\{-\frac{1}{2\delta^2}[x_{ij} - (\gamma_0 + \gamma_1 u_{ij} + \gamma_2 z_{ij})]^2\}. \tag{5.10}$$

5.3 Estimation of parameters of model (5.5) having missing observations and Measurement error

5.3.1 Estimation of parameters of model (5.5) for complete data without measurement error

The complete data likelihood with all covariates x_{ij} measured perfectly can be written as

$$\prod_{i=1}^{n} \left[\prod_{j=1}^{m_i} f(y_{ij}|x_{ij}, y_{i,j-1}, \xi_i; \theta) f(\xi_i; \sigma) \right]. \tag{5.11}$$

The log-likelihood for subject i is

$$\begin{aligned} l_i = & \sum_{j=1}^{m_i} logf(y_{ij}|x_{ij}, y_{i,j-1}; \theta) + m_i logf(\xi_i; \sigma) \\ = & \sum_{j=1}^{m_i} \Big[\big[y_{ij}(x_{ij}\beta_{01} + \xi_i) - log[1 + \exp(x_{ij}\beta_{01} + \xi_i)] \big] I_{y_{i,j-1}=0} \\ & + \big[y_{ij}(-x_{ij}\beta_{10} - \nu\xi_i) - log[1 + \exp(-x_{ij}\beta_{10} - \nu\xi_i)] \big] I_{y_{i,j-1}=1} \Big] \\ & + m_i(-\tfrac{1}{2} log 2\pi\sigma^2 - \tfrac{\xi_i}{2\sigma^2}^2). \end{aligned} \tag{5.12}$$

We can consider the random effects as missing data and use EM method here. The advantages of viewing ξ_i as missing data is that on knowing the ξ_i, all the Y_{ij}'s are independent because ξ_i can model some correlations. In addition, the M step of the EM algorithm which maximizes with respect to parameters and is similar to a standard generalized linear model computation with the values of ξ_i treated as known. The E-step: Given the observed data and current estimates of the parameters , the conditional expectation of the complete data log-likelihood for the i^{th} subject is

$$\begin{aligned} Q_i(\theta, \sigma|\theta^{(t)}, \sigma^{(t)}) = & \; E[l_i|y_i; \theta^{(t)}, \sigma^{(t)}] \\ = & \int [logf(y_i|x_i, \xi_i; \theta) + logf(\xi_i|\sigma)] \cdot f(\xi_i|y_i, x_i; \sigma^{(t)}) d\xi_i. \end{aligned} \tag{5.13}$$

The key problem in maximizing (5.13) is the integral over random effects. In some cases, this integral can have an analytical solution. However, in general, there is no closed form for it. Then we apply the Monte Carlo EM algorithm here. We need to sample a large number of ξ_i. since we have

$$f(\xi_i|y_i, x_i; \theta^{(t)}; \sigma^{(t)}) \propto f(y_i|x_i, \xi_i; \theta^{(t)}) f(\xi_i; \sigma^{(t)}), \tag{5.14}$$

Adaptive rejection sampling method can be used to produce random draws from the conditional distribution of $f(\xi_i|y_i, x_i)$. At the tth iteration, for each subject i, we generate $\xi_i^{(k)}, k = 1, \cdots, N$, from $f(\xi_i|y_i, x_i; \theta^{(t)}, \sigma^{(t)})$ and choose $\theta^{(t+1)}$ and $\sigma^{(t+1)}$ to

maximize $\sum_{i=1}^{n} 1/N \sum_{k=1}^{N} logf(y_i|x_i, \xi_i^{(k)}; \theta)$ and $\sum_{i=1}^{n} 1/N \sum_{k=1}^{N} logf(\xi_i^{(k)}; \sigma)$ respectively. If convergence is obtained, we can say $\theta^{(t+1)}$ and $\sigma^{(t+1)}$ are the maximum likelihood estimates of the parameters of model (5.5). Let $\hat{\psi} = (\theta^{(t+1)}, \sigma^{(t+1)})$.

The variance-covariance matrix of the estimates of the parameters are obtained by inverting the observed information matrix at convergence, which is

$$H = -\sum_{i=1}^{n} \frac{1}{N} \sum_{k=1}^{N} \frac{\partial^2}{\partial\psi\partial\psi'} l_i(\psi; y_i, x_i, \xi_i^{(k)}|\hat{\psi}). \tag{5.15}$$

5.3.2 Estimation of parameters of model (5.5) having missing observations

The full likelihood for incomplete data and covariates x_{ij} without measurement error is

$$\prod_{i=1}^{n}[\prod_{j=1}^{m_i} f(y_{ij}|x_{ij}, y_{i,j-1}, \xi_i, \theta) f(r_{ij}|w_{ij}, \phi) f(\xi_i, \sigma)]. \tag{5.16}$$

Then, the log-likehood for subject i is

$$\begin{aligned} l_i = & \sum_{j=1}^{m_i} logf(r_{ij}|w_{ij}, \phi) + \sum_{j=1}^{m_i} f(y_{ij}|x_{ij}, y_{i,j-1}, \xi_i, \theta) + m_i logf(\xi_i, \sigma) \\ = & \sum_{j=1}^{m_i} \left[r_{ij}(w_{ij}\phi) - log[1 + \exp(w_{ij}\phi)]\right] \\ & + \sum_{j=1}^{m_i} \Big[\left[y_{ij}(x_{ij}\beta_{01} + \gamma_1\xi_i) - log[1 + \exp(x_{ij}\beta_{01} + \gamma_1\xi_i)]\right] I_{y_{i,j-1}=0} \\ & + \left[y_{ij}(-x_{ij}\beta_{10} - \gamma_2\xi_i) - log[1 + \exp(-x_{ij}\beta_{10} - \gamma_2\xi_i)]\right] I_{y_{i,j-1}=1} \Big] \\ & + m_i(-\tfrac{1}{2} log 2\pi\sigma^2 - \tfrac{\xi_i}{2\sigma^2}^2). \end{aligned} \tag{5.17}$$

Since response y_i has missing values, we use MCEM method again. The E-Step is

$$\begin{aligned} Q_i(\theta, \sigma, \phi|\theta^{(t)}, \sigma^{(t)}, \phi^{(t)}) = & \; E\left[l_i|Y_i^{(o)}, R_i, X_i; \theta^{(t)}, \sigma^{(t)}, \phi^{(t)}\right] \\ = & \int\int [logf(r_i|w_i, \phi) + f(y_i^{(o)}, y_i^{(m)}|x_i, \xi_i, \theta) \\ & + logf(\xi_i; \sigma)] \cdot f(y_i^{(m)}, \xi_i|y_i^o, r_i, x_i; \theta^{(t)}, \sigma^{(t)}, \phi^{(t)}) dy_i^{(m)} d\xi_i, \end{aligned} \tag{5.18}$$

At the tth iteration, for each subject i, we interactively generate $\xi_i^{(k)}$ and $y^{(m)(k)}, k = 1, \cdots, N$, from $f(y_i^{(m)}|x_i, \xi_i, y_i^{(o)}, r_i; \theta^{(t)}, \sigma^{(t)})$ and $f(\xi_i|y_i, r_i, x_i; \theta^{(t)}, \sigma^{(t)}, \phi^{(t)})$ by the adaptive rejection sampling method. ξ_i is generated based on

$$f(\xi_i|y_i, r_i, x_i; \theta^{(t)}, \sigma^{(t)}, \phi^{(t)}) \propto f(y_i|x_i, \xi_i; \theta^{(t)}, \sigma^{(t)}, \phi^{(t)} f(r_i|x_i, \xi_i; \phi^{(t)}) f(\xi_i; \sigma^{(t)}). \quad (5.19)$$

Then, choose $\theta^{(t+1)}$, $\sigma^{(t+1)}$, $\phi^{(t+1)}$ to maximize $\sum_{i=1}^{n} 1/N \sum_{k=1}^{N} logf(y_i^{(o)}, y_i^{(m)}|\xi_i^{(k)}, x_i; \theta)$, $\sum_{i=1}^{n} 1/N \sum_{k=1}^{N} logf(\xi_i^{(k)}; \sigma)$ and $\sum_{i=1}^{n} 1/N \sum_{k=1}^{N} logf(r_i|\xi^{(k)}; \omega)$ respectively. If convergence is obtained, we can say $\theta^{(t+1)}$, $\sigma^{(t+1)}$, and $\phi^{(t+1)}$ are the maximum likelihood estimates of parameters of model (5.5) with missing responses. Let $\hat{\psi} = (\theta^{(t+1)}, \sigma^{(t+1)}, \phi^{(t+1)})$. Then the variance-covariance matrix of the estimates of the parameters can be obtained by (5.15).

5.3.3 Estimation of parameters of model (5.5) for complete Data with measurement error

Now we partition x_i into the error-prone covariate u_i which can only be observed through the value of v_i, and error-free covariate z_i. We also suppose $f(y_i|u_i, v_i, z_i) = f(y_i|u_i, z_i)$, which is called the non-differential error mechanism (Carroll et al., 2006, p.36).

The likelihood for the complete data with error prone covariates u_i and the error free covariate z_i is

$$\prod_{i=1}^{n} \left[\prod_{j=1}^{m_i} f(y_{ij}|u_{ij}, z_{ij}, y_{i,j-1}, \xi_i; \theta) f(u_{ij}|v_{ij}, z_{ij}; \omega, \delta) f(\xi_i; \sigma) \right]. \quad (5.20)$$

The log-likelihood for subject i is

$$\begin{aligned} l_i = & \sum_{j=1}^{m_i} logf(y_{ij}|u_{ij}, y_{i,j-1}, z_{ij}, \xi_i; \theta) + \sum_{j=1}^{m_i} logf(u_{ij}|v_{ij}, z_{ij}; \omega, \delta) + m_i logf(\xi_i; \sigma) \\ = & \sum_{j=1}^{m_i} \Big[\big[y_{ij}(u_{ij}\beta_{01u} + z_{ij}\beta_{01z} + \xi_i) - log[1 + \exp(u_{ij}\beta_{01u} + z_{ij}\beta_{01z} + \xi_i)] \big] I_{y_{i,j-1}=0} \\ & + \big[y_{ij}(-u_{ij}\beta_{10u} - z_{ij}\beta_{01z} - \nu\xi_i) - log[1 + \exp(-u_{ij}\beta_{10u} - z_{ij}\beta_{01z} - \nu\xi_i)] \big] I_{y_{i,j-1}=1} \\ & - \tfrac{1}{2} log(2\pi\delta^2) - \tfrac{1}{2\delta^2}[u_{ij} - (\gamma_0 + \gamma_1 v_{ij} + \gamma_2 z_{ij})]^2 \Big] \\ & + m_i(-\tfrac{1}{2} log 2\pi\sigma^2 - \tfrac{\xi_i}{2\sigma^2}^2). \end{aligned} \tag{5.21}$$

Since both u_{ij} and ξ_i are unobserved, they can be treated as missing values. Then, we apply MCEM method again. The E-Step is given as

$$\begin{aligned} Q_i(\theta, \omega, \delta, \sigma|\theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)}) = & \; E[l_i|y_i, v_i, z_i; \theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)}] \\ = & \int\int [logf(y_i|u_i, z_i, \xi_i; \theta) + logf(u_i|v_i, z_i; \omega, \delta) \\ & + logf(\xi_i; \sigma)] \cdot f(u_i, \xi_i|y_i, v_i, z_i; \theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)}) du_i d\xi_i, \end{aligned} \tag{5.22}$$

To use the Monte Carlo method to solve this integration problem, we need to generate a large number of samples ξ_i and u_i from $f(u_i, \xi_i|y_i, z_i, v_i; \theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)})$. Gibbs sampling technique is applied to convert a multivariate distribution sampling problem to a univariate conditional distribution problem. Based on the following

$$\begin{aligned} & f(u_i|\xi_i, y_i, v_i, z_i; \theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)}) \propto f(y_i|u_i, \xi_i, z_i; \theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)}) f(u_i|v_i, z_i; \omega^{(t)}, \delta^{(t)}) \\ & f(\xi_i|u_i, y_i, v_i, z_i; \theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)}) \propto f(y_i|u_i, \xi_i, z_i; \theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)}) f(\xi_i; \sigma^{(t)}), \end{aligned} \tag{5.23}$$

at the tth iteration, for each subject i, we interactively generate $\xi_i^{(k)}$ and $u^{(k)}, k = 1, \cdots, N$, from $f(u_i|\xi_i, y_i, v_i, z_i; \theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)})$ and $f(\xi_i|u_i, y_i, v_i, z_i; \theta^{(t)}, \omega^{(t)}, \delta^{(t)}, \sigma^{(t)})$ by adaptive rejection sampling method. Then, choose $\theta^{(t+1)}$, $\sigma^{(t+1)}$, $\omega^{(t+1)}$ and $\delta(t+1)$ to maximize $\sum_{i=1}^{n} 1/N \sum_{k=1}^{N} logf(y_i|\xi_i^{(k)}, u_i^{(k)}, z_i; \theta)$, $\sum_{i=1}^{n} 1/N \sum_{k=1}^{N} logf(\xi_i^{(k)}; \sigma)$ and $\sum_{i=1}^{n} 1/N \sum_{k=1}^{N} logf(u_i|v_i, z_i; \omega)$ respectively. If convergence is obtained, we can say $\theta^{(t+1)}$, $\sigma^{(t+1)}$, $\omega^{(t+1)}$ and $\delta^{(t+1)}$, are the maximum likelihood estimates of the parameters of model (5.5) with covariate measurement error. Let $\hat{\psi} = (\theta^{(t+1)}, \sigma^{(t+1)}, \omega^{(t+1)}, \delta^{(t+1)})$.

Then the variance-covariance matrix of the estimates of the parameters can be obtained by (5.15).

5.3.4 Estimation of parameters of model (5.5) having missing observations with measurement error

The full joint likelihood considering covariate measurement error and missing data process is the product of four conditional distributions as

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{m_i} f(y_{ij}|u_{ij},z_{ij},y_{i,j-1},\xi_i;\theta)f(r_{ij}|\xi_i,u_{ij},z_{ij},r_{i,j-1};\phi)f(u_{ij}|v_{ij},z_{ij};\omega,\delta)f(\xi_i;\sigma)\right], \tag{5.24}$$

where y_{ij} is composed of observed part $y_{ij}^{(o)}$ and missing part $y_{ij}^{(m)}$. The log likelihood contributed from subject i is

$$\begin{aligned} l_i = \;& \sum_{j=1}^{m_i} logf(r_{ij}|w_{ij};\phi) + \sum_{j=1}^{m_i} logf(y_{ij}|u_{ij},z_{ij},\xi_i;\theta) \\ & + \textstyle\sum_{j=1}^{m_i} logf(u_{ij}|v_{ij},z_{ij};\omega,\delta) + m_i logf(\xi_i;\sigma) \\ = \;& \sum_{j=1}^{m_i}\Big[r_{ij}(w_{ij}\phi) - log[1+\exp(w_{ij}\phi)]\Big] \\ & + \sum_{j=1}^{m_i}\Big[\big[y_{ij}(u_{ij}\beta_{01u}+z_{ij}\beta_{01z}+\xi_i) - log[1+\exp(u_{ij}\beta_{01u}+z_{ij}\beta_{01z}+\xi_i)]\big]I_{y_{i,j-1}=0} \\ & + \big[y_{ij}(-u_{ij}\beta_{10u}-z_{ij}\beta_{01z}-\nu\xi_i) - log[1+\exp(-u_{ij}\beta_{10u}-z_{ij}\beta_{01z}-\nu\xi_i)]\big]I_{y_{i,j-1}=1} \\ & - \tfrac{1}{2}log(2\pi\delta^2) - \tfrac{1}{2\delta^2}[u_{ij}-(\gamma_0+\gamma_1 v_{ij}+\gamma_2 z_{ij})]^2\Big] \\ & + m_i(-\tfrac{1}{2}log2\pi\sigma^2 - \tfrac{\xi_i}{2\sigma^2}^2). \end{aligned} \tag{5.25}$$

The E-step of the EM algorithm is to calculate the expected value of the complete data log-likelihood given the observed data and current parameter estimates. The

E-Step gives,

$$\begin{aligned} Q_i(\theta,\omega,\delta,\sigma,\phi|\theta^{(t)},\omega^{(t)},\delta^{(t)},\sigma^{(t)},\phi^{(t)}) = \; & E[l_i|y_i^{(o)},r_i,v_i,z_i;\theta^{(t)},\omega,\delta,\sigma^{(t)},\phi^{(t)}] \\ = \; & \int\int\int[(logf(r_i|y_i^{(o)},y_i^{(m)},u_i,z_i,\xi_i;\phi) \\ & +logf(y_i^{(o)},y_i^{(m)}|u_i,z_i,\xi_i;\theta) \\ & +logf(u_i|v_i,z_i;\omega,\delta)+logf(\xi_i;\sigma)] \\ & \cdot f(y_i^{(m)},u_i,\xi_i|y_i^{(o)},r_i,v_i,z_i;\theta^{(t)},\omega^{(t)},\delta^{(t)},\sigma^{(t)},\phi^{(t)}) \\ & dy_i^{(m)}du_i d\xi_i. \end{aligned} \tag{5.26}$$

For each subject i, at the tth iteration, the kth ($k = 1, 2, ..., N$) sample can be generated for $(y_i^{(m)(k)}, u_i^{(k)}, \xi_i^{(k)})$ by using the same method as the above scenario. For a given N, $Q_i(\theta,\omega,\delta,\sigma,\phi|\theta^{(t)},\omega^{(t)},\delta^{(t)},\sigma^{(t)},\phi^{(t)})$ can be approximated by $\frac{1}{N_i}\sum_{k=1}^{N_t} l_i$, where l_i can be calculated by replacing $(y^{(m)}, u_i, \xi_i)$ with $(y_i^{(m)(k)}, u_i^{(k)}, \xi_i^{(k)})$. In the M step, we maximize $\sum_{i=1}^{n} Q_i$ using maximum likelihood to obtain the updated estimates. We can say $\theta^{(t+1)}$, $\sigma^{(t+1)}$, $\omega^{(t+1)}$, $\delta^{(t+1)}$ and $\phi^{(s+1)}$ are the maximum likelihood estimates of the parameters of model (5.5) with missing response and covariate measurement error. Let $\hat{\psi} = (\theta^{(t+1)}, \sigma^{(t+1)}, \omega^{(t+1)}, \delta^{(t+1)}, \phi^{(s+1)})$. Then the variance-covariance matrix of the estimates of the parameters can be obtained by (5.15).

Appendix

1. The estimating equations and the elements of the observed Fisher information matrix for data under the beta-binomial Model

Estimating equations for beta binomial data are as follows.

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \left[\sum_{r=0}^{y_i-1} \frac{1}{\pi_i + r\phi} - \sum_{r=0}^{m_i-y_i-1} \frac{1}{1-\pi_i + r\phi} \right] \frac{\partial \pi_i}{\partial \beta_j} = 0,$$

$$\frac{\partial l}{\partial \phi} = \sum_{i=1}^{n} \left[\sum_{r=0}^{y_i-1} \frac{r}{\pi_i + r\phi} + \sum_{r=0}^{m_i-y_i-1} \frac{r}{1-\pi_i + r\phi} - \sum_{r=0}^{m_i-1} \frac{r}{1+r\phi} \right] = 0.$$

The elements of the observed information matrix for beta binomial data can be obtained from the elements of the second derivative matrix for beta binomial data which are given blow.

$$\frac{\partial^2 l}{\partial \beta_j \beta_s} = -\sum_{i=1}^{n} \left[\sum_{r=0}^{y_i-1} \frac{1}{(\pi_i + r\phi)^2} + \sum_{r=0}^{m_i-y_i-1} \frac{1}{(1-\pi_i + r\phi)^2} \right] \left[\frac{\partial \pi_i}{\partial \beta_j} \frac{\partial \pi_i}{\partial \beta_s} \right]$$

$$+ \left[\sum_{r=0}^{y_i-1} \frac{1}{\pi_i + r\phi} - \sum_{r=0}^{m_i-y_i-i} \frac{1}{1-\pi_i + r\phi} \right] \frac{\partial^2 \pi_i}{\partial \beta_j \beta_s},$$

$$\frac{\partial^2 l}{\partial\beta_j\partial\phi} = \sum_{i=0}^{n}\left[-\sum_{r=0}^{y_i-1}\frac{r}{(\pi_i+r\phi)^2}+\sum_{r=0}^{m_i-y_i-1}\frac{r}{(1-\pi_i+r\phi)^2}\right]\frac{\partial\pi_i}{\partial\beta_j},$$

$$\frac{\partial^2 l}{\partial\phi^2} = \sum_{i=1}^{n}\left[-\sum_{r=0}^{y_i-1}\frac{r^2}{(\pi_i+r\phi)^2}-\sum_{r=0}^{m_i-y_i-1}\frac{r^2}{(1-\pi_i+r\phi)^2}+\sum_{r=0}^{m_i-1}\frac{r^2}{(1+r\phi)^2}\right].$$

2. The Elements of the observed Fisher information matrix for data under the zero-inflated over-dispersion beta-binomial model

Define

$B_1=\prod_{r=0}^{m_i-1}(1+r\phi)$, $B_2=\prod_{r=0}^{m_i-1}(1+r\phi-\pi_i)$, $B_3=\sum_{j=0}^{m_i-1}\prod_{r=0,r\neq j}^{m_i-1}(1+r\phi-\pi_i)$, $B_4=\sum_{j=0}^{m_i-1}j\prod_{r=0,r\neq j}^{m_i-1}(1+r\phi-\pi_i)))$, $B_5=\sum_{j=0}^{m_i-1}j\prod_{r=0,r\neq j}^{m_i-1}(1+r\phi)$, $B_6=\sum_{j=0}^{m_i-1}\sum_{k=0,k\neq j}^{m_i-1}\prod_{r=0,r\neq j,r\neq k}^{m_i-1}(1+r\phi-\pi_i))$, $B_7=\sum_{j=0}^{m_i-1}\sum_{k=0,k\neq j}^{m_i-1}k\prod_{r=m,r\neq k,r\neq j}^{m_i-1}(1+r\phi-\pi_i)$, $B_8=\sum_{j=0}^{m_i-1}j\sum_{k=0,k\neq j}^{m_i-1}k\prod_{r=0,r\neq j,r\neq k}^{m_i-1}(1+r\phi)$, $A_1=\gamma+\frac{B_2}{B_1}$, $A_2=\frac{1}{B_1}-B_3$, $A_3=\frac{B_4B_1-B_2B_5}{B_1^2}$, $A_4=\frac{B_6}{B_1}$, $A_5=\frac{-B_7B_1-B_3B_5}{B_1^2}$, $A_6=\frac{-2B_5(B_4B_1-B_2B_5)+B1(B_8B_1-B_2B_8)}{B_1^3}$.

Then the elements of the observed Fisher information matrix for ZIBB model can be obtained from the elements of the second derivative matrix for ZIBB model which are written as

$$\frac{\partial^2 l_i}{\partial\beta_j\beta_s} = \left\{\left[-\left(\frac{A_2}{A_1}\right)^2+\frac{A_4}{A_1}\right]I_{\{y_i=0\}}+\left[-\sum_{r=0}^{y_i-1}\frac{1}{(\pi_i+r\phi)^2}-\sum_{r=0}^{m_i-y_i-1}\frac{1}{(1-\pi_i+r\phi)^2}\right]I_{\{y_i>0\}}\right\}\frac{\partial\pi_i}{\partial\beta_j}\frac{\partial\pi_i}{\partial\beta_s}+\left\{\frac{A_2}{A_1}I_{\{y_i=0\}}+\left[\sum_{r=0}^{y_i-1}\frac{1}{\pi_i+r\phi}-\sum_{r=0}^{m_i-y_i-i}\frac{1}{1-\pi_i+r\phi}\right]I_{\{y_i>0\}}\right\}\frac{\partial^2\pi_i}{\partial\beta_j\partial\beta_s},$$

$$\frac{\partial^2 l}{\partial\beta_j\partial\phi} = \left\{\left[-\frac{A_2A_3}{A_1^2}+\frac{A_5}{A_1}\right]I_{\{y_i=0\}}+\left[-\sum_{r=0}^{y_i-1}\frac{r}{(\pi_i+r\phi)^2}+\sum_{r=0}^{m_i-y_i-1}\frac{r}{(1-\pi_i+r\phi)^2}\right]I_{\{y_i>0\}}\right\}\frac{\partial\pi_i}{\partial\beta_j},$$

$$\frac{\partial^2 l_i}{\partial\beta_j\partial\gamma} = \left[-\frac{A_2}{A_1^2} I_{\{y_i=0\}}\right]\frac{\partial\pi_i}{\partial\beta_j},$$

$$\begin{aligned}\frac{\partial^2 l_i}{\partial\phi^2} = & \left[-\left(\frac{A_3}{A_1}\right)^2 + \frac{A_6}{A_1}\right] I_{\{y_i=0\}} \\ & + \left[-\sum_{r=0}^{y_i-1}\frac{r^2}{(\pi_i+r\phi)^2} - \sum_{r=0}^{m_i-y_i-1}\frac{r^2}{(1-\pi_i+r\phi)^2} + \sum_{r=0}^{m_i-1}\frac{r^2}{(1+r\phi)^2}\right] I_{\{y_i>0\}},\end{aligned}$$

$$\frac{\partial^2 l_i}{\partial\phi\partial\gamma} = \left(-\frac{A_3}{A_1^2}\right) I_{\{y_i=0\}},$$

$$\frac{\partial^2 l_i}{\partial\gamma^2} = \frac{1}{(1+\gamma)^2} - \frac{1}{A_1^2} I_{\{y_i=0\}},$$

where $\dfrac{\partial\pi_i}{\partial\beta_j} = x_{ij}\pi_i(1-\pi_i)$, $\dfrac{\partial\pi_i}{\partial\beta_s} = x_{is}\pi_i(1-\pi_i)$, and $\dfrac{\partial^2\pi_i}{\partial\beta_j\partial\beta_s} = x_{ij}x_{is}\pi_i(1-\pi_i)(1-2\pi_i)$.

3. Expressions for the elements of H_1 and H_2 from estimates $\hat{\psi}_1$ and $\hat{\psi}_2$

Under MNAR the observed information H_1 has the form

$$H_1 = -Q''(\psi,\alpha|\psi^{(s)},\alpha^{(s)}) = \begin{bmatrix} -Q_1''(\psi|\psi^{(s)}) & 0 \\ 0 & -Q_2''(\alpha|\alpha^{(s)}) \end{bmatrix},$$

which shows that $\hat{\psi}$ and $\hat{\alpha}$ are independent and we only need the first entry $-Q_1''(\psi|\psi^{(s)})$ in the diagonal matrix to obtain the variance-covariance matrix of $\hat{\psi}_1$.

Now, it can be seen that

$$Q_1''(\psi|\psi^{(s)}) = \sum_{i=1}^{k}\frac{\partial^2}{\partial\psi\partial\psi'} l_i(\psi|y_{o,i},x_i;\hat{\psi}_1) + \sum_{i=k+1}^{n}\sum_{y_{m,i}=0}^{m_i} w_{iy_i}^{(s)}\frac{\partial^2}{\partial\psi\partial\psi'} l_i(\psi|y_{m,i},x_i;\hat{\psi}_1),$$

where

$$w_{iy_i}^{(s)} = \frac{f(y_{m,i}|x_i;\psi^{(s)})f(r_i|x_i,y_{m,i};\alpha^{(s)})}{\sum_{y_{m,i}=0}^{m_i} f(y_{m,i}|x_i;\psi^{(s)})f(r_i|x_i,y_{m,i};\alpha^{(s)})}$$

and $l_i(\psi|y_i, x_i; \hat{\psi}_1)$ are from the ZIBB model (3.4).

The weights are calculated from model (3.3) and model (3.8). Therefore, the elements of $Q''(\psi, \alpha|\psi^{(s)}, \alpha^{(s)})$ regarding to ψ can be obtained by the expressions for the elements of the second derivative matrix for ZIBB model.

Under MAR, after deleting the model for the missing data mechanism,

$$H_2 = -Q''(\psi|\psi^{(s)}) = \quad -\sum_{i=1}^{k} \frac{\partial^2}{\partial\psi\partial\psi'} l_i(\psi|y_{o,i}, , x_i; \hat{\psi}_2) - \sum_{i=k+1}^{n} \sum_{y_{m,i}=0}^{m_i} w_{iy_i}^{(s)} \frac{\partial^2}{\partial\psi\partial\psi'} l_i(\psi|y_{m,i}, x_i; \hat{\psi}_2),$$

where

$$w_{iy_i}^{(s)} = f(y_{m,i}|x_i, \psi^{(s)}).$$

The weights are calculated from the ZIBB model (3.3) and the log-likelihood function $l_i(\psi|y_i, x_i; \hat{\psi}_1)$ is also from the ZIBB model (3.4). Then, the elements of $Q''(\psi|\psi^{(s)})$ can be also obtained by the expressions for the elements of the second derivative matrix for the ZIBB model.

4. Gibbs sampling

Gibbs sampling is the way to convert a multivariate sampling problem into a univariate sampling problem. The point of Gibbs sampling is that given a multivariate distribution it is simpler to sample from a conditional distribution. Suppose we want to obtain

$$x = (x_1, \cdots, x_p) \sim q(x_1, \cdots, x_p).$$

Denote the ith sample by $X^{(i)} = (x_1^{(i)}, \cdots, x_p^{(i)})$. We begin with some arbitrary set of initial values $((x_1^{(0)}, \cdots, x_p^{(0)})$, at interaction $i(i \geq 0)$. We sample the components in

order, starting from the first component and proceeding with the following sampling steps to get $(x_1^{(i+1)}, \cdots, x_p^{(i+1)})$.

$$
\begin{aligned}
x_1^{(i+1)} &\sim q_1(x_1 | x_2^{(i)}, \cdots, x_p^{(i)}) \\
x_2^{(i+1)} &\sim q_2(x_2 | x_1^{(i+1)}, x_3^{(i)} \cdots, x_p^{(i)}) \\
&\vdots \\
x_p^{(i+1)} &\sim q_p(x_p | x_1^{(i+1)}, \cdots, x_{p-1}^{(i+1)}).
\end{aligned}
$$

Repeat the above step k times. Geman and Geman (1987) proved the Gibbs convergence theorem that $(x_1^{(i)}, \cdots, x_p^{(i)})$ converges to $(x_1, \cdots, x_p) \sim q(x_1, \cdots, x_p)$ as $i \to \infty$. Because samples from the early iterations are not from the target posterior, it is common to discard these samples. The discarded iterations are often referred to as the "burn-in" period.

Bibliography

Adcock, R. J. (1878). A problem in least squares. *The Analyst*, *5*(2), 53–54.

Albert, P. S. (2000). A transitional model for longitudinal binary data subject to nonignorable missing data. *Biometrics*, *56*(2), 602–608.

Albert, P. S., & Follmann, D. A. (2003). A random effects transition model for longitudinal binary data with informative missingness. *Statistica Neerlandica*, *57*(1), 100–111.

Allison, P. D. (2001). *Missing data*. Thousand Oak, CA: Sage publications.

Altham, P. M. (1978). Two generalizations of the binomial distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *27*(2), 162–167.

Briggs, A., Clark, T., Wolstenholme, J., & Clarke, P. (2003). Missing.... presumed at random: Cost-analysis of incomplete data. *Health Economics*, *12*(5), 377–392.

Burr, D. (1988). On errors-in-variables in binary regression—Berkson case. *Journal of the American Statistical Association*, *83*(403), 739–743.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., & Stefanski, L. A. (2006). *Measurement error in nonlinear models: A modern perspective*. New York: Chapman and Hall/CRC.

Carroll, R. J., Spiegelman, C. H., Lan, K. G., Bailey, K. T., & Abbott, R. D. (1984). On errors-in-variables for binary regression models. *Biometrika*, *71*(1), 19–25.

Crowder, M. J. (1978). Beta-binomial Anova for proportions. *Applied Statistics*, *27*(1), 34–37.

Dean, C. B. (1992). Testing for overdispersion in Poisson and Binomial regression models. *Journal of the American Statistical Association*, *87*(418), 451–457.

Deltour, I., Richardson, S., & Hesran, J.-Y. L. (1999). Stochastic algorithms for Markov models estimation with intermittent missing data. *Biometrics*, *55*(2), 565–573.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, *39*(1), 1–38.

Deng, D., & Paul, S. (2005). Score tests for zero-inflation and over-dispersion in generalized linear models. *Statistica Sinica*, *15*, 257–276.

Deng, D., & Paul, S. R. (2000). Score tests for zero inflation in generalized linear models. *Canadian Journal of Statistics*, *28*(3), 563–570.

Donovan, A., Ridout, M., & James, D. (1994). Assessment of somaclonal variation in apple. ii. rooting ability and shoot proliferation in vitro. *Journal of Horticultural Science*, *69*(1), 115–122.

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*(3), 457–483.

Elston, R. (1977). Response to query, consultants corner. *Biometrics*, *33*, 232–233.

Fuller, W. A. (2009). *Measurement error models*. New York: John Wiley & Sons.

Geman, S., & Geman, D. (1987). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in computer vision* (pp. 564–584). Elsevier.

Gibson, G., & Austin, E. (1996). Fitting and testing spatio-temporal stochastic models with application in plant epidemiology. *Plant Pathology*, *45*(2), 172–184.

Gleser, L. J. (1981). Estimation in a multivariate "errors in variables" regression model: Large sample results. *The Annals of Statistics*, *9*(1), 24–44.

Godambe, V., & Thompson, M. E. (1989). An extension of quasi-likelihood estimation. *Journal of Statistical Planning and Inference*, *22*(2), 137–152.

Haseman, J., & Kupper, L. (1979). Analysis of dichotomous response data from certain toxicological experiments. *Biometrics*, *35*(1), 281–293.

Hogg, R. V., Tanis, E. A., & Zimmerman, D. L. (1977). *Probability and statistical inference*. New York: Macmillan.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*(411), 765–769.

Ibrahim, J. G., Chen, M. H., & Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, *88*(2), 551–564.

Ibrahim, J. G., Chen, M.-H., Lipsitz, S. R., & Herring, A. H. (2005). Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, *100*(469), 332–346.

Ibrahim, J. G., & Lipsitz, S. R. (1996). Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics*, *52*(3), 1071–1078.

Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions*. New York: John Wiley & Sons.

Kannel, W., Neaton, J., Wentworth, D. f., Thomas, H., Stamler, J., Hulley, S., & Kjelsberg, M. (1986). Overall and coronary heart disease mortality rates in

relation to major risk factors in 325,348 men screened for the MRFIT. *American Heart Journal*, *112*(4), 825–836.

Kleinman, J. C. (1973). Proportions with extraneous variance: Single and independent samples. *Journal of the American Statistical Association*, *68*(341), 46–54.

Korn, E. L., & Whittemore, A. S. (1979). Methods for analyzing panel studies of acute health effects of air pollution. *Biometrics*, *35*(4), 795–802.

Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. New York: John Wiley & Sons.

Lüning, K., Sheridan, W., Ytterborn, K. H., & Gullberg, U. (1966). The relationship between the number of implantations and the rate of intra-uterine death in mice. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *3*(5), 444–451.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, *11*(1), 59–67.

Mian, R., & Paul, S. (2016). Estimation for zero inflated negative binomial model with missing response. *Statistics in Medicine*, *35*, 5603–5624.

Nakai, M., & Ke, W. (2011). Review of the methods for handling missing data in longitudinal data analysis. *International Journal of Mathematical Analysis*, *5*(1), 1–13.

Nelder, J. A., & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, *74*(2), 221–232.

Otake, M., & Prentice, R. L. (1984). The analysis of chromosomally aberrant cells based on beta-binomial distribution. *Radiation Research*, *98*(3), 456–470.

Paul, S. (1982). Analysis of proportions of affected foetuses in teratological experiments. *Biometrics*, *38*(2), 361–370.

Prentice, R. (1982). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, *69*(2), 331–342.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). GLLAMM manual.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association*, *72*(359), 538–543.

Schafer, D. W. (1987). Covariate measurement error in generalized linear models. *Biometrika*, *74*(2), 385–391.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 257–261.

Stefanski, L. A., & Carroll, R. J. (1985). Covariate measurement error in logistic regression. *The Annals of Statistics*, *13*(4), 1335–1351.

Stubbendick, A. L., & Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, *59*(4), 1140–1150.

Troxel, A. B., Lipsitz, S. R., & Brennan, T. A. (1997). Weighted estimating equations with nonignorably missing response data. *Biometrics*, *53*(3), 857–869.

Wang, Y.-G. (1999). Estimating equations with nonignorably missing response data. *Biometrics*, *55*(3), 984–989.

Wei, G. C., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*(411), 699–704.

Weil, C. (1970). Selection of the valid number of sampling units and a consideration of their combination in toxicological studies involving reproduction, teratogenesis or carcinogenesis. *Food and Cosmetics Toxicology*, *8*(2), 177–182.

Williams, D. (1975). The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics*, *31*(4), 949–952.

Wolter, K. M., & Fuller, W. A. (1982). Estimation of nonlinear errors-in-variables models. *The Annals of Statistics*, *10*(2), 539–548.

Wu, M., & Ware, J. H. (1979). On the use of repeated measurements in regression analysis with dichotomous responses. *Biometrics*, *35*(2), 513–521.

Yang, X., Shoptaw, S., Nie, K., Liu, J., & Belin, T. R. (2007). Markov transition models for binary repeated measures with ignorable and nonignorable missing values. *Statistical Methods in Medical Research*, *16*(4), 347–364.

Zeger, S. L., & Qaqish, B. (1988). Markov regression models for time series: A quasi-likelihood approach. *Biometrics*, *44*(4), 1019–1031.

Zeng, L., & Cook, R. J. (2007). Transition models for multivariate longitudinal binary data. *Journal of the American Statistical Association*, *102*(477), 211–223.

Vita Auctoris

NAME:	Rong Luo
PLACE OF BIRTH:	Qinhuangdao, Hebei, China
YEAR OF BIRTH:	1970
EDUCATION:	Yanshan University, Qinhuangdao, China
	1988-1995 B.Sc and M.S.
	University of Toledo, Toledo, OH
	2010-2012 M.S.,
	University of Windsor, Windsor, ON
	2012-2019 PhD.